DISS. ETH NO. 30440

THE GENETIC CODE AND EVOLUTION

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES
(Dr. sc. ETH Zurich)

presented by

HANA ROZHOŇOVÁ

MSc ETH in Computational Biology and Bioinformatics
born on 22.9.1994

accepted on the recommendation of

Dr. Joshua L. Payne, examiner
Prof. Dr. Pedro Beltrao, co-examiner
Prof. Dr. Sean R. Eddy, co-examiner
Prof. Dr. Stephen J. Freeland, co-examiner

2024

SUMMARY

The genetic code is the set of rules by which organisms translate the information stored in nucleic acids into proteins. Remarkably, nearly all life forms on Earth use the same set of rules, the standard genetic code, although variant codes do exist. Besides its near-universality, another notable feature of the standard genetic code is its robustness to effects of mutations, as amino acid pairs that are one nucleotide substitution apart under the standard genetic code tend to have similar physicochemical properties.

The genetic code is mostly believed to be "frozen", incapable of evolution, because, in a sufficiently large genome, changing the meaning of even a single codon would have a prohibitively deleterious effect. However, the existence of extant non-standard genetic codes demonstrates that the genetic code is capable of evolving to a certain extent. This raises several key questions about the genetic code and evolution: (1) How did the standard genetic code evolve to its current form? (2) What evolutionary processes give rise to the extant non-standard genetic codes? (3) How does the genetic code, which forms a map between the genotype (DNA) and the phenotype (protein), influence the evolution of proteins? In this thesis, we have explored each of these questions.

Many theories have been proposed regarding the origins of the standard genetic code and the selection pressures that shaped it. Recently, it has been suggested that, in addition to their conservative nature regarding physicochemical properties of amino acids, single-nucleotide substitutions also tend to preserve the carbon and nitrogen content of amino acids under the standard genetic code. This suggests that the genetic code may have evolved under pressure for resource conservation. In Chapter 2, we scrutinize this theory and demonstrate that upon closer inspection the structure of the standard genetic code provides little evidence to support this hypothesis. More broadly, we discuss why drawing conclusions based on seemingly non-random patterns within the standard genetic code is challenging. Specifically, we highlight that such patterns may result from various confounding factors and we emphasize the importance of using an appropriate null model of alternative genetic codes for comparison to the standard genetic code.

Over the last few decades, several dozen extant alternative genetic codes have been discovered. To understand why and how these non-standard genetic codes evolve, it is essential to have a good understanding of their prevalence. In Chapter 3, we conduct a computational screen for alternative genetic codes in over 15.5 million viral genomes. Viruses, as intra-cellular parasites that mostly rely on the host cell's translation apparatus, typically evolve alternative genetic codes in response to hosts that also use alternative genetic codes. However, recently it has been suggested that some viruses also utilize the switch to an alternative genetic code as a gene regulation mechanism, controlling the production of genes needed in the final phase of the infection. In our large-scale screen, we discovered four previously unknown instances of this phenomenon in distinct groups of bacteriophages, indicating that the ability to employ a genetic code switch as a gene regulation mechanism is widespread among bacteriophages. Additionally, we identified two bacteriophage groups using alternative genetic code throughout their genomes, despite the absence of a known host using the same alternative genetic code. Finally, we found two groups of large eukaryotic viruses using an alternative genetic code likely in response to their hosts, representing the first examples of alternative genetic codes in this group of viruses.

Finally, in Chapter 4, we shift our focus from the evolution of the genetic code to its influence on evolution. Specifically, we explore the relationship between the code's robustness and the evolvability it facilitates. To do this, we examine the ruggedness of six empirical fitness landscapes under hundreds of thousands of in silico-generated rewired genetic codes. We show that robust genetic codes tend to produce smooth fitness landscapes with few adaptive peaks, which are readily accessible from throughout the sequence space in evolutionary simulations. Although the

impact of a particular genetic code on protein evolvability largely depends on the specific fitness landscape, we identify several general features of genetic codes that tend to enhance or diminish evolvability. These design principles could be used in the future to engineer codes with increased or diminished evolvability for research or industrial applications.

ZUSAMMENFASSUNG

Der genetische Code sind die Regeln, mit denen Organismen die in Nukleinsäuren gespeicherten Informationen in Proteine übersetzen. Bemerkenswerterweise verwenden fast alle Lebensformen auf der Erde dieselben Regeln, den sogenannten Standard-Code, allerdings existieren auch alternative Code Varianten. Neben seiner fast universellen Verbreitung ist ein weiteres bemerkenswertes Merkmal des Standard-Codes seine Robustheit gegenüber den Auswirkungen von Mutationen, da Aminosäuren die nur eine Nukleotidsubstitution voneinander entfernt sind, unter dem Standard-Code ähnliche physikochemische Eigenschaften aufweisen.

Der genetische Code wird größtenteils als evolutionär "eingefroren" angesehen, d.h. er ist unfähig sich zu verändern, weil eine Änderung der Bedeutung auch nur eines einzigen Codons in einem ausreichend großen Genom eine prohibitiven schädlichen Effekt hätte. Die Existenz von alternativen genetischen Codes zeigt jedoch, dass der genetische Code in gewissem Maße evolutionsfähig ist. Dies wirft mehrere Schlüsselfragen zum genetischen Code und zur Evolution auf: (1) Wie ist die heutige Form des Standard-Codes evolviert? (2) Welche evolutionären Prozesse haben zu der Entstehung der existierenden alternativen genetischen Codes beigetragen? (3) Wie beinflusst der genetische Code, der den Genotyp (DNA) zum Phänotyp (Protein) übersetzt, die Evolution von Proteinen? In dieser Dissertation haben wir jede dieser Fragen untersucht.

Es existieren viele Theorien zu den Ursprüngen des Standard-Codes und den Selektionsdrücken, die ihn geformt haben. Kürzlich wurde vorgeschlagen, dass neben der konservativen Natur in Bezug auf die physikochemischen Eigenschaften von Aminosäuren, einzelne Nukleotidsubstitutionen auch den Kohlenstoff- und Stickstoffgehalt von Aminosäuren unter dem Standard-Code erhalten. Dies deutet darauf hin, dass der genetische Code unter Druck zur Ressourcenschonung evolviert sein könnte. In Kapitel 2 überprüfen wir diese Theorie und zeigen, dass bei genauerer Betrachtung die Struktur des Standard-Codes wenig Beweise zur Unterstützung dieser Hypothese liefert. Allgemeiner diskutieren wir, warum es schwierig ist, Schlussfolgerungen basierend auf scheinbar nicht zufälligen Mustern innerhalb des Standard-Codes zu ziehen. Insbesondere betonen wir, dass solche Muster durch verschiedene Störfaktoren entstehen können und dass ein geeignetes Nullmodell für alternative genetische Codes unabdingbar ist um sinnvolle Vergleiche mit dem Standard-Code durchzuführen.

In den letzten Jahrzehnten wurden mehrere Dutzend alternative genetische Codes entdeckt. Um zu verstehen, warum und wie diese alternativen genetischen Codes evolvieren, ist es entscheidend, ein gutes Verständnis ihrer Verbreitung zu haben. In Kapitel 3 führen wir einen computergestützten Screen nach alternativen genetischen Codes in über 15,5 Millionen Virusgenomen durch. Viren, als intrazelluläre Parasiten, die größtenteils auf die Translations-Machinerie der Wirtszelle angewiesen sind, evolvieren typischerweise alternative genetische Codes als Reaktion auf ihre Wirte, welche ebenfalls alternative genetische Codes verwenden. Es wurde jedoch kürzlich vorgeschlagen, dass einige Viren den Wechsel zu einem alternativen genetischen Code auch als Mechanismus der Genregulation nutzen und so die Produktion von Genen steuern, die in der Endphase der Infektion benötigt werden. In unserem Screen entdeckten wir vier zuvor unbekannte Fälle dieses Phänomens in verschiedenen Gruppen von Bakteriophagen, was darauf hindeutet, dass Codewechsel zur Genregulationsmechanismus bei Bakteriophagen weit verbreitet sind. Darüber hinaus identifizierten wir zwei Bakteriophagengruppen, die in ihrem gesamten Genom einen alternativen genetischen Code verwenden, obwohl kein Wirt mit demselben alternativen genetischen Code bekannt ist. Desweiteren fanden wir zwei Gruppen großer eukaryotischer Viren, die wahrscheinlich als Reaktion auf ihre Wirte einen alternativen genetischen Code verwenden und damit die ersten Beispiele für alternative genetische Codes in dieser Gruppe von Viren darstellen.

Abschließend wechseln wir in Kapitel 4 unseren Fokus von der Evolution des genetischen Codes zum Einfluss des genetischen Codes auf Evolution generell. Insbesondere untersuchen wir die Beziehung zwischen der Robustheit des Codes und der Evolvierbarkeit, die er ermöglicht. Dazu untersuchen wir die Schroffheit von sechs empirischen Fitnesslandschaften unter Hunderttausenden von in silico-generierten genetischen Codes. Wir zeigen, dass robuste genetische Codes tendenziell glatte Fitnesslandschaften mit wenigen adaptiven Gipfeln produzieren, welche in evolutionären Simulationen leicht aus dem gesamten Sequenzraum erreichbar sind. Obwohl der Einfluss eines bestimmten genetischen Codes auf Protein-Evolvierbarkeit weitgehend von der spezifischen Fitnesslandschaft abhängt, identifizieren wir mehrere allgemeine Merkmale von genetischen Codes, die Evolvierbarkeit fördern oder verringern. Diese Merkmale könnten in Zukunft verwendet werden, um Codes mit erhöhter oder verringerter Evolvierbarkeit für Forschungs- oder industrielle Anwendungen zu entwickeln.

CONTENTS

1

INTRODUCTION

But if thought corrupts language, language can also corrupt thought.
— George Orwell

The genetic code is the set of rules by which organisms translate information encoded in nucleic acids into proteins. In particular, each of the 64 possible triplets of nucleotides (codons) corresponds to an amino acid or a stop signal. Nearly all life forms on Earth, from bacteria to humans, use the same genetic code – the standard genetic code (Fig. 1.1) – though variant codes do exist [1].

The genetic code is implemented by the translation system which is universally conserved in all extant life forms [3]. The process involves two distinct steps (Fig. 1.2). First, an aminoacyl tRNA synthetase (aaRS) specifically recognizes and covalently binds a trnasfer RNA (tRNA) with the corresponding amino acid. Most cells posses a set of 20 different types of aaRSs, one for each amino acid [4]. In the ribosome, the anticodon of the aminoacylated tRNA then basepairs with the complementary codon on the messenger RNA (mRNA) and the amino acid is transferred onto the growing peptide chain. Thus, there is no need for a direct recognition of the codons by the amino acids. Other components of the translation apparatus include release factors, which are proteins that terminate translation by specifically recognizing stop codons, and enzymes responsible for modifying and/or editing the tRNAs to alter their recognition by the synthetase and/or the ribosome [5], [6].

Immediately after the deciphering of the standard genetic code in the 1960s it became evident that the assignment of amino acids to codons is not random [7], [8]. Codons are organized into synonymous blocks, where alterations in the third position often do not change the amino acid meaning. Moreover, even these synonymous blocks do not seem to be allocated in a random way – for example, all codons with U in the second position encode hydrophobic amino acids (Fig. 1.1). This intuitive notion of non-randomness in the standard genetic code was confirmed by two seminal papers by Haig and Hurst [9] and Freeland and Hurst [10]. These studies introduced a quantitative measure of error tolerance of the code, defined as the expected squared change in polar requirement (a measure of hydrophilicity; Fig. 1.1) upon a single-nucleotide substitution. By comparing the error tolerance of the standard genetic code with those of a large number of alternative, in silico-generated genetic codes, Freeland and Hurst demonstrated that the standard genetic code exhibits an exceptional, "one in a million" degree of error tolerance to the effects of point mutations. Subsequent research has also revealed the code's propensity to preserve physicochemical properties of amino acids upon frameshift mutations [11], [12] and its ability to enrich for adaptive mutations [13].

The origins and evolution of the genetic code have long captivated scientists [7], [8], [14], perhaps because there is a certain tension between its remarkable properties and its perceived static nature in the face of evolutionary dynamics. As articulated by Vetsigian et al. [15], "The genetic code could well be optimized to a greater extent than anything else in biology and yet is generally regarded as the biological element least capable of evolving." Indeed, it is widely

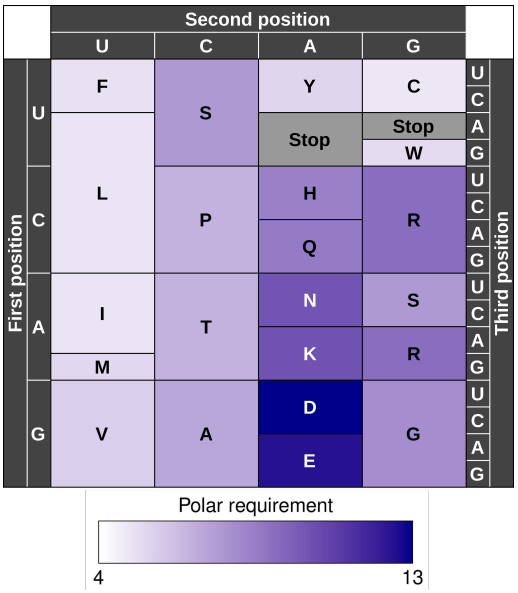


FIGURE 1.1: The standard genetic code. The codons are colored based on the polar requirement of the corresponding amino acid [2], a measure of hydrophilicity. The first, second, and third codon positions correspond to the letters on the left, top, and right of the table, respectively.

accepted that once the codon table is established, any attempt to alter even a single codon in a sufficiently large genome will have prohibitively deleterious effect. How, then, did life converge upon an error-minimizing, seemingly highly optimized, genetic code, if modifying it is impossible, yet variation is a necessary condition of adaptation? How and why do extant alternative genetic codes evolve? And what are the implications of the error tolerance exhibited by the standard genetic code for the adaptive potential of organisms? Despite decades of research, these and other questions concerning the relationship between the genetic code and evolution remain unanswered. In this thesis, we aim to address some of the gaps in our understanding of this relationship.

1.1 ORIGINS OF THE STANDARD GENETIC CODE

The three main theories attempting to explain the structure of the standard genetic code are the stereochemical, coevolution (metabolic), and error minimization theory.

The stereochemical theory proposes that, unlike in the modern translation machinery (Fig. 1.2), genetic code evolution was driven by a direct physicochemical affinity between amino acids and the corresponding codons or anticodons [16], [17]. It is true that, for some amino acids, short RNA sequences selected from random mixtures by amino acid binding are significantly enriched for the cognate triples (codons in some cases and anticodons in others) [17]–[20]. However, these findings primarily involve amino acids believed to be late additions to the genetic code, such as arginine, histidine, and tryptophan, which limits their explanatory power regarding the code's origins.

The coevolution theory posits that the genetic code was shaped by metabolic relationships between amino acids [21]. According to this theory, the code evolved from an ancestral version that only contained simple amino acids through a process of subdivision. Specifically, as new

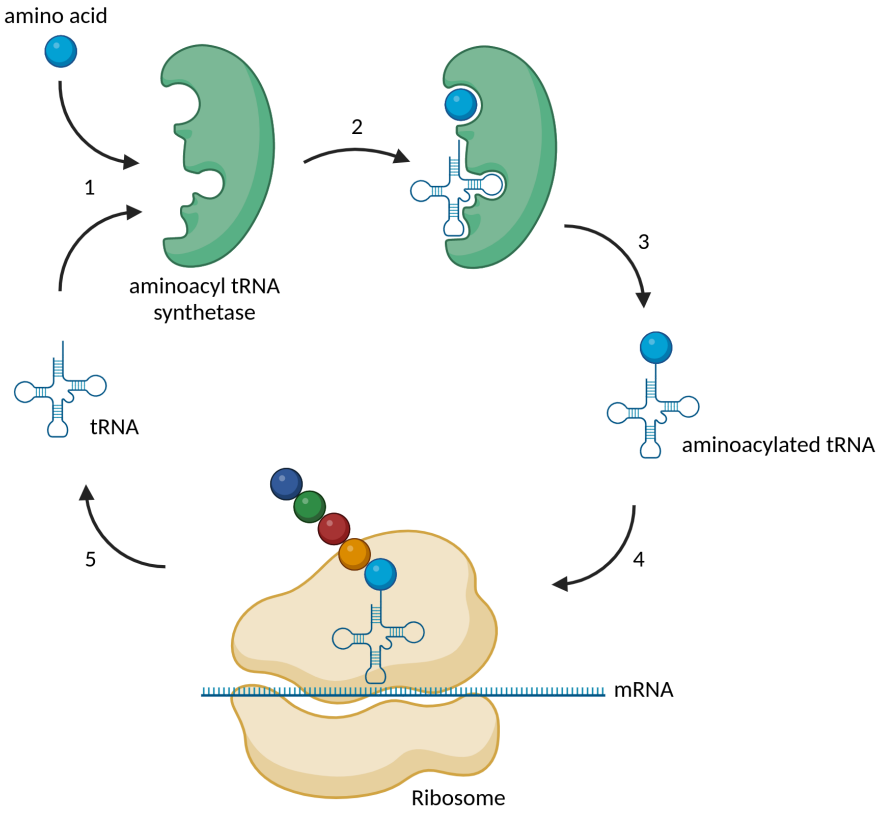
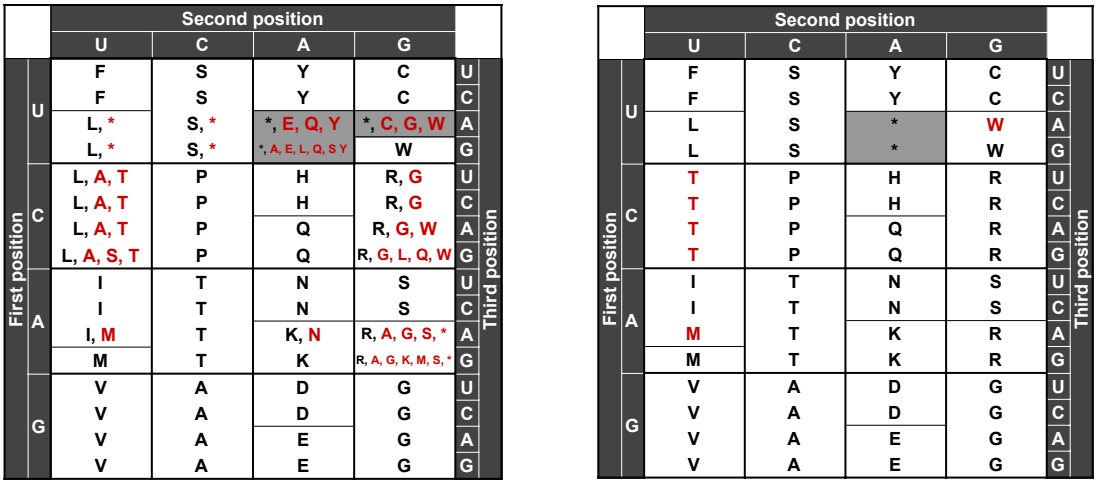


FIGURE 1.2: A schematic representation of the translation process. An aminoacyl tRNA synthetase specifically recognizes a unique amino acid and the corresponding tRNA(s) (1) and charges the tRNA with the amino acid (2). The aminoacylated tRNA (3) then travels to the ribosome (4) where its anticodon pairs with the codon on the mRNA. The amino acid is transferred onto the growing peptide chain and the tRNA is released and recycled (5). Created with Biorender.com.

metabolic pathways emerged, the codon block associated with the precursor amino acid in the ancestral code would be split into two, with the product amino acid inheriting some of the codons of the precursor amino acid. The evolution of the code is thus thought to be driven by positive selection for a broader amino acid repertoire, facilitating a wider range of protein functions. The theory's validity has been supported by a number of computational studies [21]–[23]. Additionally, it has been argued that tRNA-dependent amino acid transformation pathways, where a tRNA is misaminoacylated and then biosynthetically transformed (a typical example would be charging of bacterial Gln-type tRNAs with glutamate, which is then, while covalently bound onto the tRNA, converted to glutamine) [4], are "molecular fossils" of the coevolution mechanism [24]. On the other hand, criticism has been raised regarding the selection of the precursor-product pairs and the statistical significance of the relationship between code structure and biosynthetic pathways has been questioned [25], [26].

Unlike the metabolic pathways relationship, posited to shape the structure of the genetic code under the coevolution theory, the code's robustness in terms of amino acid hydrophobicity and, to a lesser extent, molecular volume has been extensively validated through numerous studies [9], [10], [27]–[34]. Given the extreme level of optimization observed in the standard genetic code, one might naturally conclude that the genetic code evolved under selection pressure to minimize the effects of mutations. However, an alternative explanation has been suggested in 2008 by Stephen Massey [35], who showed that the emergence of error tolerance is possible even under conditions of neutrality. Similarly to the coevolution theory, he assumes that the code evolved through subdivision of larger codon blocks. He further assumes that when a codon block splits, the amino acid chosen to occupy the newly created codon block is physicochemically similar to the original amino acid. Such tendency to allocate neighboring codon blocks to physicochemically similar amino acids might arise e.g. because the code expansion would likely require a duplication of an aaRS (or its proto-life alternative) and the newly duplicated aaRS is expected to recognize a physicochemically similar amino acid and a tRNA with related anticodon [36]. A related explanation suggests that in the early stages following a block subdivision, when the genome has still not fully adjusted to the new genetic code, only alterations allocating the newly created codon blocks to physicochemically similar amino acids would not adversely affect the organism's fitness [37].

Regardless of the exact mechanism by which the genetic code came into existence, a fundamental question remains: Why is the standard genetic code universal across all life forms? It is hard to believe that throughout the evolutionary history of the code, alternative codes never emerged. After all, none of the theories described above provide deterministic predictions of the assignment of codons to amino acids. Why, then, has only one genetic code survived to this day? The answer to this question may lie in the importance of horizontal gene transfer, which likely played a significant part in early evolution and continues to shape microbial evolution to this day [38], [39]. Even small alterations to the genetic code can have adverse effect on horizontal gene transfer [40], [41]. Through mathematical modeling, it has been demonstrated that the presence of horizontal gene transfer in a population using multiple genetic codes leads to the convergence upon a single, universal code with a structure and error tolerance akin to that of the standard genetic code [15], [42].



(a) (b)

FIGURE 1.3: Extant alternative genetic codes. (a) All currently known codon reassignments. The standard genetic code meaning of each codon is in black, the known codon reassignments are in red. (b) The yeast mitochondrial code. Changes relative to the standard genetic code are highlighted in red.

1.2 ALTERNATIVE GENETIC CODES AND THEIR EVOLUTION

As discussed above, the standard genetic code (Fig. 1.1) is used by nearly all organisms on Earth. However, over the last five decades, variations of the standard genetic code have been discovered in several dozen lineages of bacteria, archaea, eukaryotes, and mitochondrial genomes (Fig. 1.3a) [1], [43], [44]. All known genetic code alterations are limited in scope, typically involving changes in the meaning of only a couple of codons relative to the standard genetic code. The most extensive known alteration is six codon reassignments in the code used by yeast mitochondria (Fig. 1.3b) [45]–[47]. The vast majority of the alternative genetic codes preserve the amino acid alphabet of the standard 20 proteinogenic amino acids. The only two additional amino acids encoded by some genetic codes are selenocysteine, incorporated in a context-dependent manner in response to the UGA stop codon in all three domains of life [48], [49], and pyrrolysine, encoded by the UAG stop codon in a small number of archaea and bacteria [50]–[52].

How and why do alternative genetic codes evolve? On a molecular level, codon reassignment can happen through several possible mechanisms. In theory, mutations in any component of the translation apparatus — the ribosome, tRNAs, aaRSs, or release factors — could lead to changes in the genetic code. In practice, most alternative genetic codes arise from changes to tRNAs and/or release factors. For tRNAs, changes in the anticodon can lead to the tRNA recognizing a different codon. Such changes can result from mutations in the tRNA gene sequence [53], [54], alterations in tRNA base modifications that affect the repertoire of codons the tRNA can efficiently recognize [55]–[57], or changes in intron splicing [58]. Additionally, mutations in the identity elements — the sequence features the aaRSs use to recognize their cognant tRNAs — can lead to changes in tRNA aminoacylation [59], [60]. When a stop codon is lost or gained, alterations in release factors, proteins that specifically recognize stop codons and terminate translation, are

typically involved. These alterations include release factor loss [61]–[64] or modifications to their specificity [65], [66].

Altering the translational machinery alone is not sufficient for a successful codon reassignment though; the genome must also adapt by removing the codon from positions where the new meaning is intolerable. It was the necessity of this extensive removal of the reassigned codon from the genome that led Francis Crick to postulate the famous "frozen accident" hypothesis [8]. Today, there are three main models that attempt to explain how this evolutionary barrier might be crossed. In the "codon capture" model [67], the codon is driven to extinction by forces unrelated to its translation, such as a bias in genome GC content. This extinction allows the codon to be "captured" by a different amino acid. The "unassigned codon" model [68], [69] proposes that the codon's meaning is lost, for example due to the loss of the corresponding tRNA, leaving the codon unassigned and declining in frequency until new translational machinery evolves to translate it. In the "ambiguous intermediate" model [70], the new decoding machinery is acquired before the old meaning is lost, leading to a period when the codon is translated ambiguously. Selection then removes the codon from sites where the new meaning is not tolerated and introduces it at sites where the new meaning is adaptive. Eventually, the original translational machinery is lost and the new meaning fixed. Of course, these three models are not mutually exclusive and it is likely that in reality codon reassignments result from a combination of these mechanisms. For example, if a codon becomes rare due to an extreme GC content bias, it is easier for it to be reassigned via an ambiguous intermediate.

Knowing how, the last question remains: Why do codon reassignments evolve? It is certainly possible that the emergence of alternative genetic codes is a neutral process driven by genetic drift, as suggested by the "codon capture" model [67]. However, several reasons have been suggested for why codon reassignments might be adaptive. It has been argued that many extant alternative genetic codes exhibit an increased level of error tolerance compared to the standard genetic code [71], [72], although whether this is true or not depends on the exact definition of error tolerance [29]. The "genome reduction" hypothesis suggests that codon reassignments are driven by selection to minimize genome size by decreasing the number of components needed for translation (e.g., tRNAs) [73], [74]. Codon reassignment might improve protein function by rescuing a deleterious nonsense or missense mutation [75] or by introducing new biochemical capabilities. For example, the "22nd" amino acid pyrrolysine enables methanogenic archaea and bacteria to metabolize methylamine in anaerobic conditions [51]. Ambiguous translation, the proposed intermediate stage of the "ambiguous intermediate" model, has been observed to be adaptive in stress conditions [76]. Lastly, alternative genetic codes may provide protection from viruses and selfish genetic elements [40], [41], [77], [78].

1.3 INFLUENCE OF THE GENETIC CODE ON EVOLUTION

So far, we have discussed the past and ongoing evolution of the genetic code. However, the genetic code is more than a passive translation table; it is an active participant in the evolutionary process. While mutations occur at the nucleotide level, selection often operates at the level of proteins. The genetic code acts as a crucial link between these two levels, forming what is known as a genotype-phenotype map [79] and specifying which protein sequences are "near" each other in

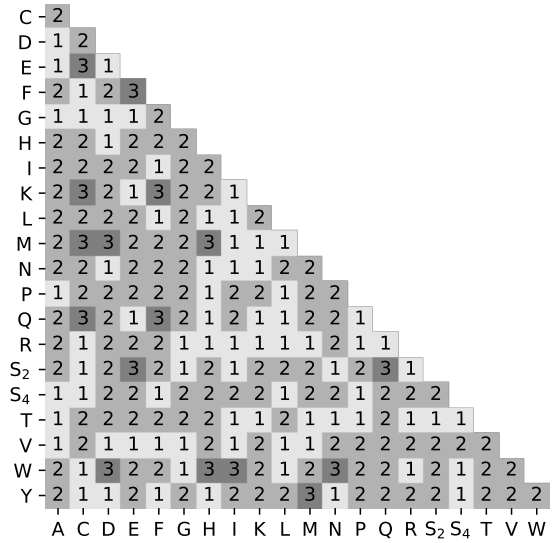


FIGURE 1.4: The minimum number of single-nucleotide substitutions needed to transform a codon of one amino acid into a codon of another amino acid, for all amino acid pairs. S_2 and S_4 denote the two disjoint sets of serine codons (Fig. 1.1): $S_2 = \{\text{AGU}, \text{AGC}\}$, $S_4 = \{\text{UCU}, \text{UCC}, \text{UCA}, \text{UCG}\}$.

the sequence space [80], [81]. For example, single-nucleotide substitutions are the most prevalent type of mutation. However, the triplet structure of the genetic code means that some codon pairs are multiple mutational steps away, restricting which amino acid substitutions are accessible via single-nucleotide mutations (Fig. 1.4).

We have both empirical and theoretical evidence for the effect of the genetic code on adaptive evolution. Looking at a large number of sequence alignments, Gonnet et al. [82] observed that the structure of the genetic code affects which mutations are accepted during evolution. Another line of evidence comes from comparing the evolutionary potentials of synonymous encodings of the same amino acid or the same gene. Codon volatility, defined as the proportion of codon's point-mutation neighbors that encode different amino acids, has been suggested as a means to detect selection by identifying genes with higher or lower volatility than expected by chance [83]. However, this method of selection detection remains controversial [84]–[87]. In any case, volatile codons appear more often than their less volatile counterparts in the complementarity-determining regions of antibodies [88], which are responsible for specific antigen binding and must evolve rapidly in response to antigen presence during the process of affinity maturation [89]. Additionally, synonymous encodings of the same gene, which differ in terms of which protein sequences are accessible from them via single-nucleotide substitutions, have been shown to evolve along different evolutionary paths [90]–[92].

The obvious limitation of these empirical results is that, since almost all protein evolution on Earth occurs under the standard genetic code, the empirical evidence also pertains to the standard genetic code. Consequently, it is challenging to predict how evolution would unfold under hypothetical genetic codes that differed significantly from the standard code. Laboratory evolution using synthetic non-standard genetic codes presents an exciting direction for future research [40], [93], however, most current insights into how evolution might proceed under

alternative genetic codes come from computational studies [13], [94], [95]. These studies simulate evolutionary processes to compare the rate of adaptation under the standard genetic code with in silico-generated alternative genetic codes. They demonstrate that error-tolerant genetic codes, such as the standard genetic code, enhance the rate of adaptive evolution [94], enrich for adaptive amino acid changes [13], and enable exploration of a larger fraction of the space of functional variants [95].

While the past and ongoing evolution of the standard genetic code has fascinated scientists for decades, the influence of the genetic code on evolution itself remains much less studied. Despite some empirical and theoretical evidence demonstrating the genetic code's role in shaping adaptive evolution, there is still much to uncover. Future research, particularly in the area of synthetic non-standard genetic codes, holds promise for deepening our understanding of this topic.

1.4 THESIS OUTLINE

In this thesis, I explore the relationship between the genetic code and evolution from three different perspectives.

In Chapter 2, I examine a recently proposed theory regarding the selection forces that have shaped the structure of the standard genetic code, namely the hypothesis that the standard genetic evolved under selection for resource conservation. Through a rigorous computational analysis, I demonstrate that there is little evidence to support this hypothesis.

In Chapter 3, I investigate the recent evolution of the genetic code. I conduct a large-scale computational screen to identify alternative genetic codes in viruses, discovering eight previously unknown genetic code alterations. I discuss the potential functional implications of these genetic code alterations.

In Chapter 4, I focus on the impact of the genetic code on adaptive evolution of proteins. In particular, I study whether robust, error-tolerant codes enhance or hinder evolvability. I show that, perhaps counterintuitively, robust genetic codes promote evolvability by rendering smooth adaptive landscapes on which evolution can easily find pathways to adaptation. Additionally, I discuss these findings in the context of synthetic non-standard genetic codes which could be engineered in the lab using the existing technology and identify several genetic codes that could potentially increase evolvability beyond the level provided by the standard genetic code.

2

LITTLE EVIDENCE THE STANDARD GENETIC CODE IS OPTIMIZED FOR RESOURCE CONSERVATION

Published as: Hana Rozhoňová & Joshua L. Payne (2021). Little evidence the standard genetic code is optimized for resource conservation. *Molecular Biology and Evolution*, 38, 5127–5133. https://doi.org/10.1093/molbev/msab236

Authors' contributions: H.R. and J.L.P. designed research; H.R. performed research; H.R. and J.L.P. analyzed data; and H.R. and J.L.P. wrote the paper.

Language is the archives of history.

— Ralph Waldo Emerson

2.1 ABSTRACT

Selection for resource conservation can shape the coding sequences of organisms living in nutrient-limited environments. Recently, it was proposed that selection for resource conservation, specifically for nitrogen and carbon content, has also shaped the structure of the standard genetic code, such that the missense mutations the code allows tend to cause small increases in the number of nitrogen and carbon atoms in amino acids. Moreover, it was proposed that this optimization is not confounded by known optimizations of the standard genetic code, such as for polar requirement or hydropathy. We challenge these claims. We show the proposed optimization for nitrogen conservation is highly sensitive to choice of null model and the proposed optimization for carbon conservation is confounded by the known conservative nature of the standard genetic code with respect to the molecular volume of amino acids. There is therefore little evidence the standard genetic code is optimized for resource conservation. We discuss our findings in the context of null models of the standard genetic code.

2.2 INTRODUCTION

The standard genetic code exhibits numerous optimizations [96]. For example, the missense and frameshift mutations allowed by the standard genetic code tend to preserve key physicochemical properties of amino acids, such as polar requirement, hydropathy, and to a lesser extent, molecular volume [9], [11], [12], [28], [97], [98]. Recently, an additional optimization was proposed, namely for resource conservation [99]. Motivated by the observation that selection for resource conservation can shape the coding sequences of organisms living in nutrient-limited environments [100]–[107], it was hypothesized that selection for resource conservation has also shaped the structure of the standard genetic code, such that the missense mutations the standard genetic code allows tend to cause small increases in the number of nitrogen and carbon atoms in amino acids. Moreover, it was hypothesized that this optimization is not confounded by known optimizations of the standard genetic code, such as for polar requirement or hydropathy.

Optimizations in the standard genetic code are typically identified using one of two approaches [108]. In the "engineering approach", the standard genetic code is compared to codes found by analytical methods or heuristic search algorithms that minimize some objective function, such as the mean absolute change in polar requirement caused by missense mutations [109]–[113]. In the "statistical approach", the standard genetic code is compared to a large number of randomized codes [9], [114]. The hypothesis that the standard genetic code is optimized for resource conservation was tested using the statistical approach, specifically by quantifying the expected random mutation cost (ERMC) of missense mutations allowed by the standard genetic code, measured in units such as the number of nitrogen or carbon atoms or the absolute change in polar requirement or hydropathy of amino acids, and comparing this cost to those incurred by one million randomized codes [99].

Method	Preserves the number of codons per amino acid	Preserves the exact block structure of the standard genetic code
Quartet shuffling	yes	no[a]
Amino acid permutation	no	yes
Restricted amino acid permutation	no[b]	yes
N-Block Shuffler	yes	no[a]
Codon Shuffler	yes	no
AAAGALOC Shuffler	no	no
Random expansion	no	yes
Ambiguity reduction 1	no	yes
Ambiguity reduction 2	no	yes
2-1-3 model	no	yes

TABLE 2.1: Two key properties of the randomized genetic codes generated with the ten different methods used in this study (Methods).

[a] The randomized codes have a block structure, but it is different from that of the standard genetic code.
[b] The number of codons per amino acid is allowed to change by at most two, relative to the standard genetic code.

Because the space of randomized codes is so large ($\approx 1.5 \cdot 10^{84}$) [115], it is necessary to draw comparisons with a sample from this space, which can then be used as a null model [96]. There are many methods for generating randomized codes, and different methods can generate codes with different properties [116]. For example, a method known as quartet shuffling generates randomized codes by shuffling quartet blocks — the blocks of four codons that share the first two bases (e.g., AAA, AAC, AAG, AAU) [114], [115]. This method, which was used to test the hypothesis that the standard genetic code is optimized for resource conservation [99], generates randomized codes that preserve two key properties of the standard genetic code, namely the number of codons per amino acid and the degeneracy of the third base (e.g., the three codons for isoleucine always have the same first and second base as the codon for methionine). In contrast, a method known as amino acid permutation generates randomized codes by permuting the twenty standard amino acids amongst the synonymous codon blocks [9]. This method, which is most commonly used in the field [9], [10], [27], [29]–[33], [95], [115], [117], [118], generates randomized codes that preserve a different key property of the standard genetic code, namely the structure of the synonymous codons blocks. Importantly, in these randomized codes, the number of codons per amino acid can change drastically relative to the standard genetic code, because the permutation of amino acids amongst the synonymous blocks is random. These two methods therefore generate randomized codes with substantially different structural properties.

Here, we test the hypothesis that the standard genetic code is optimized for resource conservation with respect to nitrogen and carbon content by drawing comparisons with randomized codes generated using ten different methods, including quartet shuffling and amino acid permutation (Tab. 2.1; Methods). With respect to nitrogen, we find consistent statistical support for resource conservation using only one of the ten methods. With respect to carbon, we find consistent statistical support for resource conservation across the ten methods, but show that this optimization is confounded by the known conservative nature of the standard genetic code with respect to the molecular volume of amino acids [28].

2.3 RESULTS

2.3.1 *Computing the expected random mutation cost*

We compute the expected random mutation cost (ERMC) as

$$\text{ERMC} = \sum_{v,v' \in \mathcal{V}, v \neq v'} \text{Freq}(v) \cdot \text{Prob}(v \to v') \cdot \text{Cost}(v \to v'), \tag{2.1}$$

where $\mathcal{V}$ is the set of all 64 codons, $\text{Freq}(v)$ is the frequency of codon v (Supp. Section S2.1), $\text{Prob}(v \to v')$ is the probability of mutation from codon v to v' given a genetic code (standard or randomized) and mutation rates (e.g., based on a transition:transversion ratio), and $\text{Cost}(v \to v')$ is the cost of mutating codon v to v' [99]. For resource conservation, the cost is defined as the increase in the number of nitrogen or carbon atoms in the amino acid encoded by codon v' relative to the amino acid encoded by v, whereas for amino acid properties such as polar requirement, hydropathy, and molecular volume, it is defined as the absolute difference in the respective property of the amino acid encoded by codon v' and the amino acid encoded by v [99].

We use three sets of codon frequencies and mutation rates to compute the ERMC of a genetic code [99] (Methods), namely

1. *Baseline parameters*: All codon frequencies are equal and mutation rates are based on a transition:transversion ratio of 1:2.
2. *Ocean parameters*: Codon frequencies and mutation rates are derived from marine metagenomics samples [99].
3. *Diverse species parameters*: Codon frequencies are derived from 39 species [119] and mutation rates are based on 11 transition:transversion ratios ranging from 1:5 to 5:1. In total, this set includes 429 combinations of codon frequencies and mutation rates [99].

For each set, we determine the statistical significance of the ERMC of the standard genetic code by computing an empirical p-value, which is the fraction of 1 million randomized genetic codes that have an ERMC that is less than or equal to that of the standard genetic code. We compute this empirical p-value separately for each of the ten methods for generating randomized codes. For the *diverse species parameters*, we correct the p-values for testing multiple hypotheses [120].

The codon frequencies for the *ocean* and *diverse species parameters*, the *ocean* mutation rates, and the raw and corrected p-values for all tests can be found in Supplementary Data online [121].

2.3.2 *Nitrogen conservation is highly sensitive to choice of null model*

We find consistent statistical support for nitrogen conservation in the standard genetic code using only one of the ten methods for generating randomized codes, namely the codon shuffler [115] (*Baseline parameters*: $p = 1.00 \cdot 10^{-6}$; *Ocean parameters*: $p = 3.00 \cdot 10^{-6}$; *Diverse species parameters*: $p \leq 0.016$). For the remaining nine methods, nitrogen conservation is never consistently statistically significant across all tested parameters, as illustrated for amino acid permutation in Fig. 2.1 (*Baseline parameters*: $p = 0.485$; *Ocean parameters*: $p = 0.115$; *Diverse species parameters*: $p \geq 0.573$).

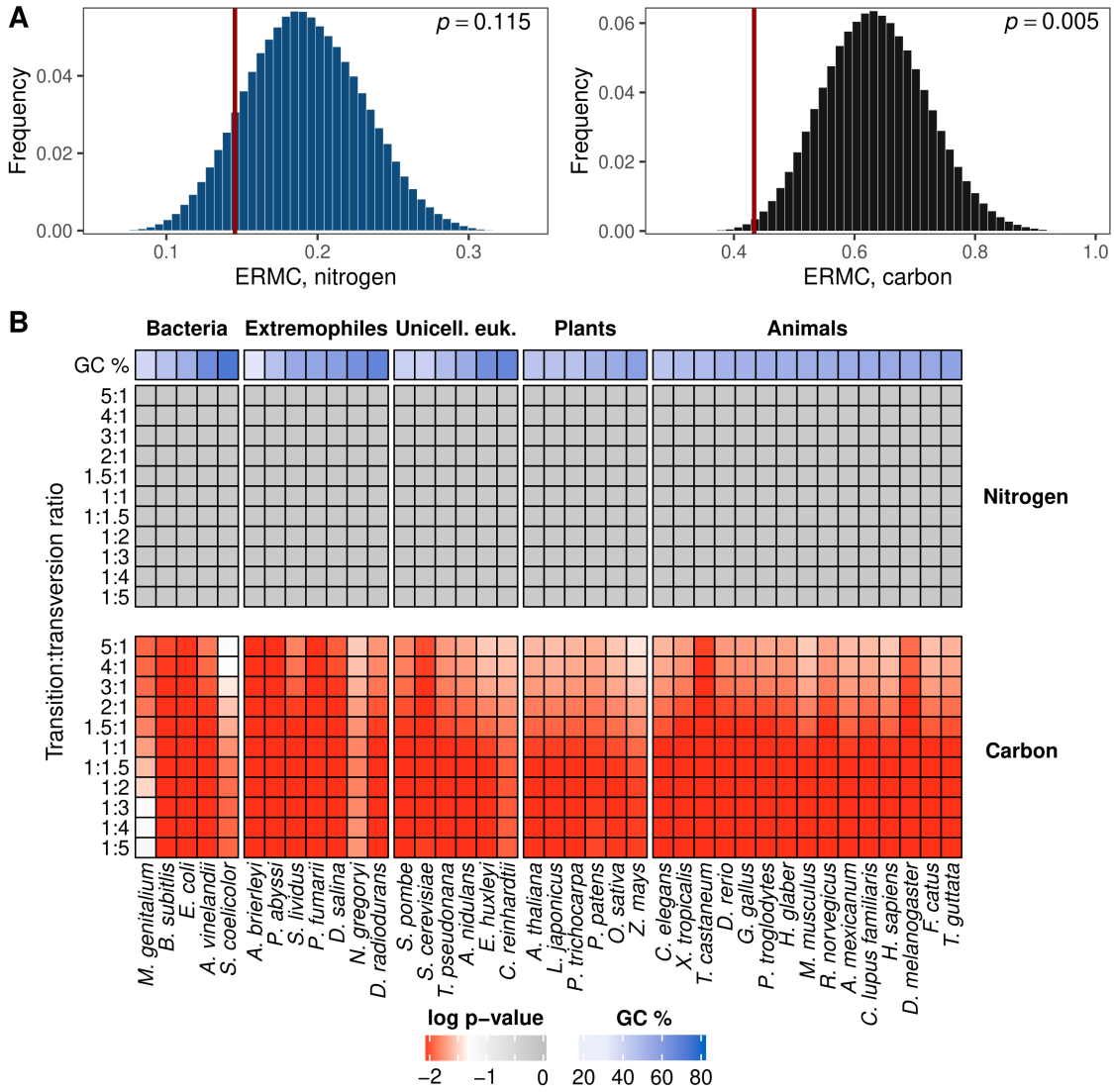


FIGURE 2.1: Nitrogen conservation is highly sensitive to choice of null model. (A) Histograms of the ERMC for nitrogen (blue) and carbon (black) in 1 million randomized codes generated by amino acid permutation. The vertical red line corresponds to the standard genetic code. Codon frequencies and mutation rates are from the *ocean parameters*. (B) P-values of the ERMC for nitrogen (top) and carbon (bottom) of the standard genetic code, relative to 1 million randomized codes generated by amino acid permutation, using the *diverse species parameters*. Shades of gray correspond to statistically insignificant p-values ($p > 0.05$; darker = less significant) and shades of red to statistically significant p-values ($p \leq 0.05$; darker = more significant). The p-values were adjusted using Benjamini-Hochberg correction for multiple testing. Organisms in each group are ordered based on the GC content of their coding sequences. Unicell. euk. = unicellular eukaryotes.

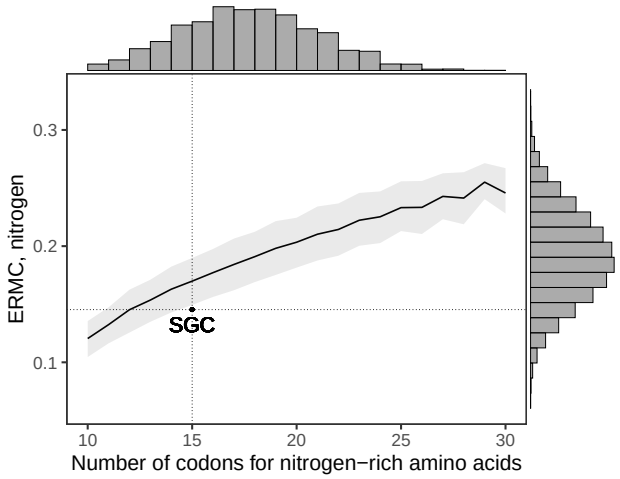


FIGURE 2.2: The ERMC for nitrogen is correlated with the number of codons for nitrogen-rich amino acids. The black line shows the mean, and the shaded area shows the 25th to the 75th quantile, of the ERMC for nitrogen in relation to the number of codons for nitrogen-rich amino acids in 1 million randomized codes generated by amino acid permutation. The point and dotted lines correspond to the standard genetic code. Histograms of the number of codons for nitrogen-rich amino acids and the ERMC for nitrogen are shown on the top and on the right of the main panel, respectively. The ERMC for nitrogen was computed using the *ocean parameters*.

Surprisingly, even for randomized codes generated by quartet shuffling, nitrogen conservation is not statistically significant for the *diverse species parameters* ($p \geq 0.316$), although it is for the *baseline parameters* ($p = 0.023$) and *ocean parameters* ($p = 0.034$).

What explains the qualitative difference between the results obtained using these different methods? The key is that the randomized codes generated by both the codon shuffler and quartet shuffling maintain the number of codons per amino acid. In the standard genetic code, the codons of the six nitrogen-rich amino acids (i.e., those with at least one nitrogen atom in their side chain: histidine, lysine, asparagine, glutamine, arginine, and tryptophan) are clustered in the codon table (Supp. Section S2.2), such that a point mutation to a codon of a nitrogen-rich amino acid leads with probability 48.9% to a codon of the same or different nitrogen-rich amino acid. Because such clustering is highly unlikely in randomized codes that maintain the number of codons per amino acid, these almost always have a higher ERMC for nitrogen than the standard genetic code, thus rendering nitrogen conservation statistically significant. In contrast, if the number of codons per amino acid is allowed to change, many randomized codes have a lower ERMC for nitrogen than the standard genetic code. The reason is that the ERMC for nitrogen is strongly correlated with the number of codons for nitrogen-rich amino acids in these randomized genetic codes (Pearson's correlation 0.567, $p < 2.2 \cdot 10^{-16}$ for codes generated by amino acid permutation; Fig. 2.2) and this number is often smaller than in the standard genetic code, thus rendering nitrogen conservation statistically insignificant.

2.3.3 *Carbon conservation is confounded by the molecular volume of amino acids*

We find consistent statistical support for carbon conservation in the standard genetic code across the ten methods for generating randomized codes (*Baseline parameters*: $p < 0.05$ for 10 of 10

methods; *Ocean parameters*: $p < 0.05$ for 9 of 10 methods; *Diverse species parameters*: median $p < 0.05$ for 9 of 10 methods; Fig. 2.1). However, we hypothesize that carbon conservation is confounded by molecular volume [122], because the molecular volume of an amino acid is strongly correlated with its number of carbon atoms (Pearson's correlation 0.906, $p = 3.97 \cdot 10^{-8}$; Fig. 2.3A) and the changes caused by missense mutations to amino acids' molecular volume and number of carbon atoms are therefore strongly correlated (Pearson's correlation 0.813, $p < 2.2 \cdot 10^{-16}$; Fig. 2.3B). We test this hypothesis using a hierarchical model [99]. Specifically, for each of the ten methods for generating randomized codes, we examine the subset of randomized codes that have an ERMC that is less than or equal to that of the standard genetic code for molecular volume and test whether the standard genetic code is also optimized for carbon conservation relative to this subset. It is not (*Baseline parameters*: $p > 0.05$ for 10 of 10 methods; *Ocean parameters*: $p > 0.05$ for 10 of 10 methods; *Diverse species parameters*: minimum $p > 0.05$ for 9 of 10 methods), as illustrated in Fig. 2.3C for randomized codes generated by amino acid permutation (*Baseline parameters*: $p = 0.139$; *Ocean parameters*: $p = 0.125$; *Diverse species parameters*: $p > 0.190$). Thus, carbon conservation is confounded by the known conservative nature of the standard genetic code with respect to the molecular volume of amino acids [28].

2.4 DISCUSSION

We found that the proposed optimization of the standard genetic code for nitrogen conservation [99] is highly sensitive to choice of null model. Specifically, we only found statistical support for nitrogen conservation when using null models that preserve the number of codons per amino acid from the standard genetic code. Choosing an appropriate null model to test for optimizations in the standard genetic code is challenging, because different null models preserve different key properties of the standard genetic code, while perturbing others. Which key properties should be preserved and which should be perturbed? This is a difficult question. On the one hand, null models that preserve the number of codons per amino acid can be justified by correlations between the number of codons per amino acid and the molecular weight of amino acids [123]–[125] as well as the frequency of amino acids in the proteome [30]. However, these correlations are far from perfect (Pearson's $R = -0.45$, $p = 0.046$ and $R = 0.67$, $p = 0.001$, respectively) and modest changes in the number of codons per amino acid are commonly observed in extant non-standard genetic codes [1]. On the other hand, null models that preserve the structure of the synonymous codon blocks can be justified by the mode of interaction between mRNA, tRNA, and the ribosome [126], [127], which results in the third 'wobble position' of codons [128]. However, extant non-standard genetic codes often have synonymous codon blocks that differ from those of the standard genetic code, demonstrating that the exact block structure of the standard genetic code is not the only possible structure [1]. Given these challenges, a sensible way forward is to use a diversity of null models when testing for optimizations in the standard genetic code [116] and to refrain from reporting optimizations that only find statistical support from a small number of these null models.

Indeed, we found such broad statistical support across a diversity of null models for the proposed optimization for carbon conservation [99], but we also found that this optimization is confounded by the known conservative nature of the standard genetic code with respect

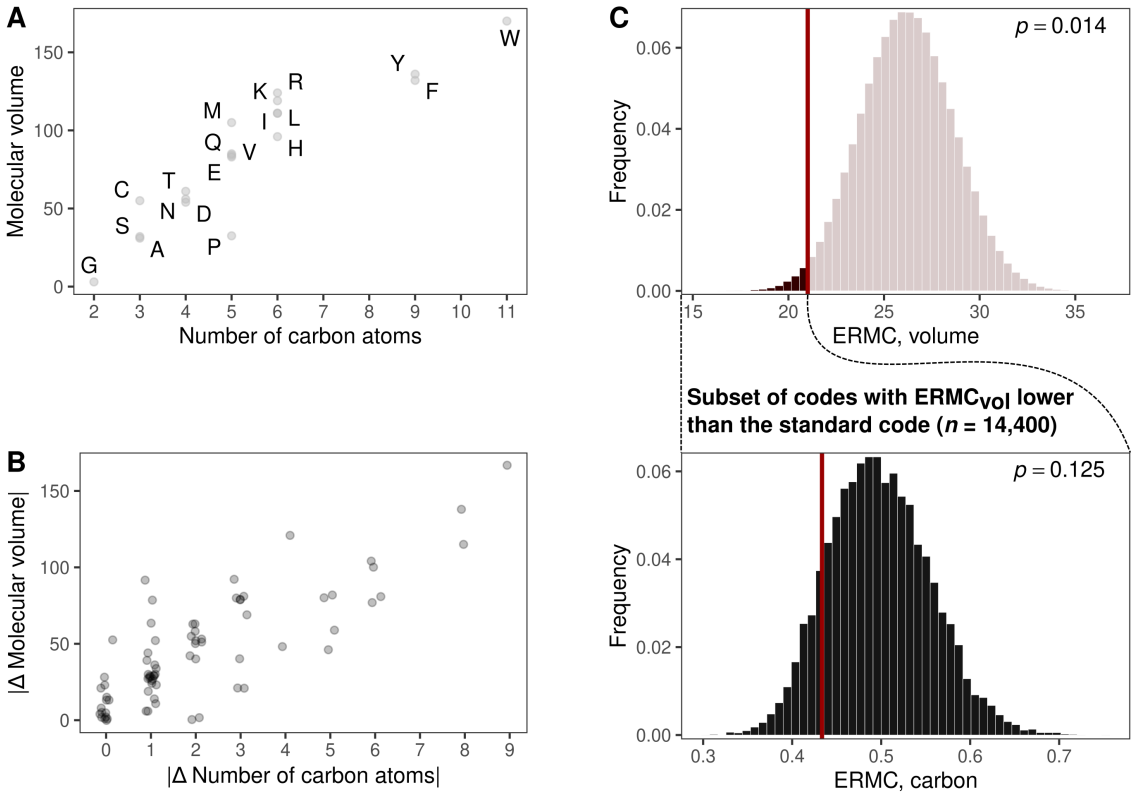


FIGURE 2.3: Carbon conservation is confounded by the molecular volume of amino acids. (A) Scatter plot of the number of carbon atoms and the molecular volume of the twenty proteinogenic amino acids. (B) Scatter plot of the absolute change in the number of carbon atoms and the absolute change in molecular volume for the 75 amino acid pairs that are connected by a missense mutation in the standard genetic code. Jitter applied in the x-axis for visualization. (C) Histograms of the ERMC for (top) the molecular volume of amino acids in 1 million randomized codes generated by amino acid permutation and (bottom) for carbon in the subset of 14,400 randomized codes that have an ERMC for molecular volume that is less than or equal to that of the standard genetic code. The ERMC was computed using the *ocean parameters*.

to molecular volume [28]. This highlights another challenge in choosing an appropriate null model to test for optimizations in the standard genetic code: Most null models are agnostic to the evolutionary history of the standard genetic code, which can give the false impression that optimizations are the product of selection rather than a byproduct of the physical processes of gene duplication and mutation ([35], [129]; but, see [130]). Carbon conservation is a case in point. While there are several non-mutually exclusive hypotheses for how the genetic code evolved [14], one hypothesis suggests that in the early stages of code evolution, amino acids were recognized by pockets in the tertiary structure of proto-tRNAs and that the expansion of the code proceeded via duplication and mutation of these proto-tRNAs [36]. Because a recently duplicated proto-tRNA would likely recognize an amino acid with similar molecular volume to that recognized by its parent proto-tRNA [35], gene duplication and mutation would naturally result in a clustering of codons for amino acids with similar molecular volumes in the codon table, as present in the standard genetic code. As carbon is the main building block of all proteinogenic amino acids, the proposed optimization for carbon conservation follows naturally, without needing to evoke selection for resource conservation. More importantly, no matter which model of genetic code evolution one considers, the endpoint is always the standard genetic code, which is conservative with respect to molecular volume [28]. This will always confound carbon conservation.

Finally, we note that if in nutrient-limited environments it is costly for missense mutations to increase the number of nitrogen or carbon atoms in amino acids, then it should be beneficial for missense mutations to decrease the number of nitrogen or carbon atoms in amino acids. This simple fact makes it difficult to justify the ERMC as a measure of the cost of missense mutations, because it only accounts for increases. Indeed, contemporaneous work to ours shows that the standard genetic code is not optimized for resource conservation when the ERMC is modified to also account for decreases in the number of nitrogen or carbon atoms [131]. Taken together, our analyses strongly suggest that the standard genetic code is not optimized for resource conservation.

2.5 METHODS

Generating randomized codes

We use ten diverse methods for generating randomized genetic codes:

- *Quartet shuffling* [99], [114], [115]: This method randomly permutes the so-called quartet blocks — quartets of codons that share the first and second nucleotide (e.g., AAA, AAC, AAG, and AAU). Moreover, it requires that the two sets of stop codons are separated by a single transition mutation [99].
- *Amino acid permutation* [9]: This method randomly permutes the twenty amino acids amongst the synonymous codon blocks.
- *Restricted amino acid permutation*: This method is the same as amino acid permutation, except the set of permutations is restricted to those in which the number of codons per amino acid changes by at most two, relative to the standard genetic code.

- *N-Block Shuffler* [115]: This method randomly permutes amino acids amongst synonymous codon blocks of the same size.
- *Codon Shuffler* [115]: This method randomly assigns codons to amino acids, ensuring that the number of codons per amino acid is the same as in the standard genetic code.
- *AAAGALOC Shuffler* [115]: This method generates randomized codes at random, only requiring that each of the twenty amino acids has at least one codon.
- *Random expansion* [35]: This method generates randomized codes by the sequential assignment of amino acids to codon blocks. The first amino acid is chosen at random and is assigned to a randomly chosen codon block. Subsequent amino acids are also chosen at random, but with probability proportional to the inverse of their Grantham distance [122] to the last amino acid added to the code[1]. This amino acid is assigned at random to one of the codon blocks that have at least one codon within one mutation from a codon for the last amino acid added to the code. This method is called 'Model 1' in ref. [35].
- *Ambiguity reduction 1* [35]: Similarly to *Random expansion*, this method generates randomized codes by sequential assignment of amino acids to codon blocks. Initially, there are only two large codon blocks, each consisting of 32 codons. The amino acids encoded by these two codon blocks are chosen at random. These large codon blocks are then subsequently split into smaller blocks, in a specified order. In each step, a codon block is split into two smaller blocks; one of them encodes the amino acid originally encoded by the block before splitting, while the second one is assigned a randomly chosen amino acid, chosen with probability proportional to the inverse of its Grantham distance to the original amino acid. This method is called 'Model 2a' in ref. [35].
- *Ambiguity reduction 2* [35]: This method works the same as *Ambiguity reduction 1*, only the order in which the codon blocks are split is different. It is called 'Model 2b' in ref. [35].
- *2-1-3 model* [35]: Similar to *Ambiguity reduction 1* and *Ambiguity reduction 2*, this method generates randomized codes by splitting the codon blocks according to the 2-1-3 model, i.e., assuming that the codon positions acquired coding potential in the order of the second, first, and then third position [132]. The first four amino acids are fixed (V, A, D, G, encoded by codons with U, C, A, and G, respectively, in the second codon position). This method is called 'Model 3' in ref. [35].

For all methods except quartet shuffling, the positions of the stop codons were fixed as in the standard genetic code.

Codon frequencies and mutation rates

For the *ocean parameters*, we obtained the codon frequencies and mutation rates directly from David Zeevi [99]. For the *diverse species parameters*, we obtained the codon frequencies from the source code of ref. [99].

1 In ref. [35], the subsequent amino acid is chosen with equal probability from all amino acids with Grantham distance to the parent amino acid smaller than a chosen threshold; we use an alternative approach to avoid parameterizing the models.

Correcting for multiple testing

For the *diverse species parameters*, the p-values were corrected using the `p.adjust` function in R, with the `method` parameter equal to `"BH"` (Benjamini-Hochberg correction for multiple testing [120]).

Data availability

Code used in this study is freely available at https://github.com/parizkh/resource-conservation-in-genetic-code.

ACKNOWLEDGEMENTS

We thank Sinisa Bratulic and David M. McCandlish for discussions and feedback on this manuscript.

2.6 SUPPLEMENTARY MATERIAL

S2.1 *Treatment of codon frequencies in computing ERMC*

The ERMC accounts for codon frequencies (Eq. 1), which are known for many species [119]. As originally formulated, the ERMC does this accounting by setting $\mathrm{Freq}(v)$ to the median frequency of all codons for amino acid a, rather than using the actual frequency of codon v from the species of interest [99]. That is, it assumes that all codons for the same amino acid have the same frequency. What is more, this median frequency is always calculated using the standard genetic code, even when measuring the ERMC of randomized codes. That is, the frequency of codons encoding amino acid a is always the same as in the standard genetic code, no matter which codons map to amino acid a in a randomized code. While an argument could be made for assuming a fixed abundance of amino acids in the proteome, we do not see the biological relevance of extending this argument to assuming fixed codon frequencies. In making it, valuable information on codon frequencies is lost, which is significant because codon usage patterns differ widely among species [133], also due to selection for resource conservation [99]. We therefore do not follow this treatment of codon frequencies in our main analyses, and instead use codon frequencies directly as they were reported for each species of interest.

Nonetheless, to facilitate direct comparison to previous work [99], we repeat all of our analyses using this treatment of codon frequencies. Defining the abundance of amino acid a as the sum of the frequencies of all codons encoding a in the standard genetic code, we consider two alternative definitions of the frequency $\mathrm{Freq}(v)$ of codon v in a randomized code:

1. *Median frequencies*: $\mathrm{Freq}(v)$ is defined as the median frequency of all codons encoding the same amino acid as v in the standard genetic code. This treatment is identical to how the codon frequencies were defined in ref. [99]. However, if the number of codons per amino acid is allowed to change in the randomized codes, this method does not preserve the abundance of each amino acid from the standard genetic code in the randomized codes.

2. *Mean frequencies*: $\mathrm{Freq}(v)$ is defined as the abundance of the amino acid encoded by v divided by the number of codons encoding the amino acid in the randomized code. This treatment is different from how the codon frequencies were defined by [99], because it takes the mean rather than the median. However, if the number of codons per amino acid is allowed to change in the randomized codes, this method preserves the abundance of each amino acid from the standard genetic code in the randomized codes.

Our results are qualitatively unchanged by these modified codon frequencies, with two exceptions: (1) Nitrogen conservation becomes statistically significant for one additional method for generating randomized codes, namely the N-Block shuffler, for both alternative treatments of codon frequencies (*Median frequencies*: *Baseline parameters*: $p = 0.014$; *Ocean parameters*: $p = 3.60 \cdot 10^{-3}$; *Diverse species parameters*: median $p = 0.031$; *Mean frequencies*: *Baseline parameters*: $p = 0.014$; *Ocean parameters*: $p = 2.94 \cdot 10^{-3}$; *Diverse species parameters*: $p \leq 0.047$). (2) Carbon conservation becomes statistically significant after controlling for the confounding factor of molecular volume for 4 of 10 methods for generating randomized codes, but only when using the *ocean parameters* and

mean frequencies. The raw and corrected p-values for all tests can be found in Supplementary Data online [121].

S2.2 *Arrangement of codons for nitrogen-rich amino acids*

It was proposed that nitrogen conservation is driven by structural principles of the standard genetic code, namely by the arrangement of synonymous codon blocks for nitrogen-rich amino acids in the codon table [99]. The standard genetic code exhibits a "square arrangement", in which codons of five of the six nitrogen-rich amino acids (histidine, lysine, asparagine, glutamine, and arginine, but not tryptophan) span only two nucleotides in their first and second positions (codons CAN, CGN, AAN and AGR, with N denoting any nucleotide and R denoting A or G). Using quartet shuffling to generate randomized codes, it was argued that this "square arrangement" (Fig. S2.1A, middle) amplifies nitrogen conservation relative to other arrangements, such as a "diagonal arrangement", in which the four codon blocks span all four nucleotides in both their first and second positions (Fig. S2.1A, right) [99]. We uncover an additional arrangement that is even more conducive to nitrogen conservation, which we call a "line arrangement" (Fig. S2.1A, left). In it, all codons for nitrogen-rich amino acids share the same nucleotide in the first or the second position (i.e., these codons occupy one row or one column of the codon table). When generating randomized codes with quartet shuffling, those that exhibit a "line arrangement" tend to have a significantly lower ERMC for nitrogen than randomized codes with other arrangements (Fig. S2.1B; e.g., $p < 2.2 \cdot 10^{-16}$, "line arrangement" vs. "square arrangement", Welch two sample t-test). The reason is the "line arrangement" maximizes the number of missense mutations between codons of nitrogen-rich amino acids, and hence minimizes the number of missense mutations from codons for non-nitrogen-rich amino acids to codons for nitrogen-rich amino acids.

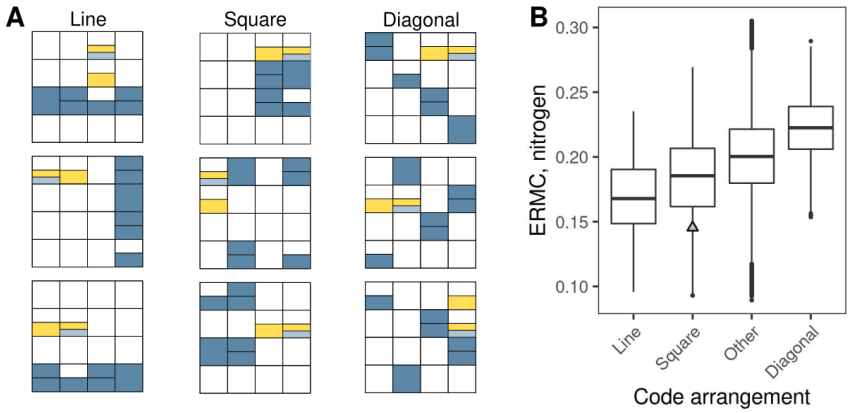


FIGURE S2.1: The arrangement of codons for nitrogen-rich amino acids in the codon table influences the ERMC for nitrogen.
A – Schematic depiction of three possible configurations of randomized genetic codes generated by quartet shuffling. Yellow boxes denote stop codons, dark blue boxes denote codons for the amino acids histidine, lysine, asparagine, glutamine, and arginine, and the light blue box is for the tryptophan codon. The topmost square-arranged code is the standard genetic code.
B – The ERMC for nitrogen shown in relation to the different arrangements of codons for nitrogen-rich amino acids in the codon tables of randomized genetic codes, computed using 1 million randomized codes generated by quartet shuffling. The grey triangle corresponds to the SGC. Codes denoted as "other" do not fall into any of the other three categories ("line", "square" or "diagonal"). The number of codes falling into each of the categories is 5,072 for "line", 20,983 for "square", 961,979 for "other", and 11,966 for "diagonal". The ERMC for nitrogen is computed using the *ocean parameters*.

3

EIGHT NEW ALTERNATIVE GENETIC CODES IN BACTERIOPHAGE AND LARGE EUKARYOTIC VIRUSES

Currently in preparation as: Hana Rozhoňová & Sean R. Eddy (2024). Eight new alternative genetic codes in bacteriophage and large eukaryotic viruses.

Authors' contributions: H.R. and S.R.E. designed research; H.R. performed research; H.R. analyzed data; and H.R. wrote the paper.

Kolik jazyků umíš, tolikrát jsi člověkem.

— Czech proverb

3.1 ABSTRACT

Viruses typically utilize the genetic code of their hosts. However, some viruses have evolved alternative genetic codes, either as an adaptation to hosts using alternative genetic codes or independently as a regulatory mechanism controlling the production of structural and lysis-related genes. In this study, we employed Codetta, a computational method for predicting the genetic code from a nucleotide sequence, to screen over 15.5 million viral contigs for alternative genetic codes. This analysis revealed five distinct genetic code variations, four of which were previously unknown in viruses, appearing in eight different clades of both prokaryotic and eukaryotic viruses. Among these were four new instances of genetic code-mediated late-phase switches in diverse clades of bacteriophages, indicating that the ability to use a genetic code change to control the progress of the viral lifecycle is widespread among bacteriophages. Additionally, we identified two groups of large eukaryotic viruses that likely use alternative genetic codes as an adaptation to their host's genetic code. Unexpectedly, we also observed two clades of bacteriophages that appear to use an alternative genetic code throughout their genomes, even though no known bacteria use these genetic codes. This suggests that these viruses either infect an as-yet-undiscovered host using an alternative genetic code, or have a unique mechanism enabling them to encode most of their genes using a genetic code different from their host.

3.2 INTRODUCTION

The genetic information encoded in the nucleic acids is translated to proteins using the genetic code, where each triplet of nucleotides (codon) specifies an amino acid or a termination signal. Nearly all organisms on Earth utilize the same variant of the genetic code, known as the standard genetic code. Yet, variations of this genetic code have been identified in several dozen lineages across all three domains of life, as well as in numerous organellar genomes [1], [43], [59].

Viruses are intracellular parasites that rely on the host cell's translation machinery for their replication. Consequently, the genetic code used by viruses typically mirrors that of their host. Indeed, there are instances where viruses infecting hosts with alternative genetic codes have evolved to use the same alternative genetic code. For example, the *Mycoplasmatales* and *Entomoplasmatales* orders of bacteria utilize a genetic code where the UGA stop codon encodes tryptophan, and viruses infecting these hosts use the same genetic code [77]. Similar genetic code coevolution has been observed in RNA viruses infecting fungal mitochondria, which also translate UGA to tryptophan [77], [134], and in bacteriophages infecting a clade of uncultivated *Bacillales* that decode the arginine codon AGG as methionine [59]. Alternatively, a virus can adapt to a host with an alternative genetic code by avoiding the use of the reassigned codon, as seen in RNA viruses infecting fungal hosts that translate the leucine codon CUG as serine [135].

In recent years, it has become apparent that some viruses employ alternative genetic codes also independent of their host. Several bacteriophage lineages have recoded the UAG or UGA stop codon, specifically UAG to glutamine and UGA to tryptophan or glycine, even though their

presumed hosts use the standard genetic code [136]–[141]. These alternative genetic codes are used only in specific parts of the genome, with the remaining parts encoded using the host's genetic code. It is believed that this genetic code change serves as a gene regulation mechanism [140]: Genes needed in the early phase of the infection utilize the host's genetic code, allowing them to be translated by the host's translation machinery. Late-phase genes, such as those encoding structural or lysis-related proteins, contain in-frame stop codons and cannot be effeciently translated under the host's genetic code. Only once the phage-encoded molecules responsible for the genetic code switch, such as a suppressor tRNA, an aminoacyl tRNA synthetase, and/or a release factor, are expressed can these late-phase proteins be produced. Thus, the genetic code change functions as a late-phase switch. Notably, early-phase proteins can still be produced even after the genetic code change, as they are "dual-coded" – the corresponding open reading frames remain nearly identical under both genetic codes as they rarely use the reassigned codon [142].

To date, such host-independent alternative genetic codes have mostly been reported in two groups of large bacteriophages, the CrAss-like and Lak phages [137]–[139]. Two systematic screens have attempted to identify alternative genetic codes also outside these groups [140], [142]. Both studies searched collections of metagenomic contigs for sequences where gene density significantly increases under a UAG- or UGA-recoded code compared to the standard genetic code. By doing so, they were able to identify instances of stop codon recoding in diverse clades of bacteriophages. Their approach, however, has two major limitations: First, it can only detect a reassignment of a stop codon to an amino acid, not changes in the amino acid meaning of sense codons. Second, it does not infer the amino acid meaning of the recoded codon, which must be determined manually by inspecting multiple-sequence alignments.

Here, we overcome these two limitations by using Codetta, a computational method that predicts amino acid meaning of each codon from the nucleotide sequence [59]. Codetta first aligns the input sequence to a collection of profiles of conserved protein domains. By doing so, it obtains, for each of the 64 possible codons, a collection of profile positions aligned to the codon. Each profile position is represented by a vector of 20 probabilities, describing the amino acid frequencies observed at a given position in a given domain. Codetta then aggregates the information from the individual profile positions to infer the most likely amino acid meaning.

We applied Codetta to a collection of more than 15.5 million viral contigs of both prokaryotic and eukaryotic viruses from the IMG/VR database [143]. This analysis revealed eight previously unknown events of genetic code change. These genetic code alterations appear in various clades of bacteriophages as well as large eukaryotic viruses, indicating that genetic code alterations are even more widespread among viruses than previously believed.

3.3 RESULTS

3.3.1 *Computational screen for alternative genetic codes in viral genomes*

We used Codetta to predict the genetic code for 15,677,623 viral genomes from the IMG/VR database [143], the largest publicly available collection of viral sequences. The database contains sequences of both prokaryotic and eukaryotic viruses. The length of the sequences ranges between

165 and 2,473,870 bp, with the mean of 10,189 bp, and many of the genomes are estimated to be incomplete (Supp. Fig. S3.1).

In total, the Codetta analysis identified 5,593 putative codon reassignments, occurring in 5,585 sequences (Supp. Fig. S3.2). Majority of these (5,077/5593, 90.8%) are predictions of a stop codon gaining an amino acid meaning. To prioritize high-confidence genetic code changes and to distinguish alternative genetic codes that are virus-specific from those that are host adaptations, we gathered additional information on these proposed reassignments: (1) We checked for the presence of the translational apparatus (tRNA, aminoacyl tRNA synthetase, and/or release factor genes) that could enact the reassignment. (2) In cases where a stop codon was predicted to encode an amino acid, we compared gene density, defined as the total length of genes divided by the length of the genome, assuming the standard genetic code and assuming the alternative genetic code. If the stop codon reassignment is real, we expect to see a significant (at least 5-10%) increase in gene density [140]. (3) We checked functions of the genes using the reassigned codon(s). If these genes have a certain function in common (e.g., they are mostly structural and lysis genes), this is evidence towards a gene regulatory function of the recoding. (4) We also checked whether the genes using the reassigned codons are clustered in the genomes. If the reassigned codon is used only in part of the genome, it is again evidence towards the genetic code change serving a regulatory function. To do this, we first identified genes containing the reassigned codon(s). Then, we computed the number of gene pairs such that the two genes are neighbors in the genome and one of them uses the reassigned codon(s) and the other not. In the end, we compared this number with a distribution generated by randomly shuffling the order of the genes in the genome and checked whether it was significantly lower than expected by chance. (5) Finally, we checked the phylogenetic extent of the proposed change by generating a gene-sharing network of the recoded genomes, together with 4,433 RefSeq *Caudoviricetes* genomes and 145 RefSeq *Nucleocytoviricota* genomes using vConTACT2 [144]. In the network, two genomes are connected with an edge if the number of homologous genes they share is significantly higher than expected by chance and genomes that are connected with an edge are expected to be closely related. Fig. 3.1 shows selected connected components (denoted Component A to E) of the resulting gene-sharing network.

Majority of the 5,593 codon reassignments predicted by Codetta are the three genetic code changes previously described in viruses: reassignment of the UAG stop codon to glutamine (2,643 occurrences), reassignment of the UGA stop codon to glycine (1,387 occurrences), and reassignment of the UGA codon to tryptophan (820 occurrences). In addition to these, we identified 4 previously unknown genetic code changes in bacteriophages (*Caudoviricetes*): reassignment of the UAG stop codon to alanine, glutamate, and serine, and reassignment of the UAA and UAG stop codons to alanine. One of these, the UAG-Glu reassignment, is present in three different viral clades (Fig. 3.1). Of these six instances of genetic code change, four appear to serve as a late-phase switch and two are used throughout the genome, but they do not seem to be host-mediated. Additionally, we also identified two different alternative codes – the reassignments of UAG to glutamine and of UAG and UAA to glutamine – in *Megaviricetes*, eukaryotic large viruses. Both of these genetic code changes seem to be an adaptation to the host genetic code. Below, we describe each of the newly discovered genetic code changes in more detail.

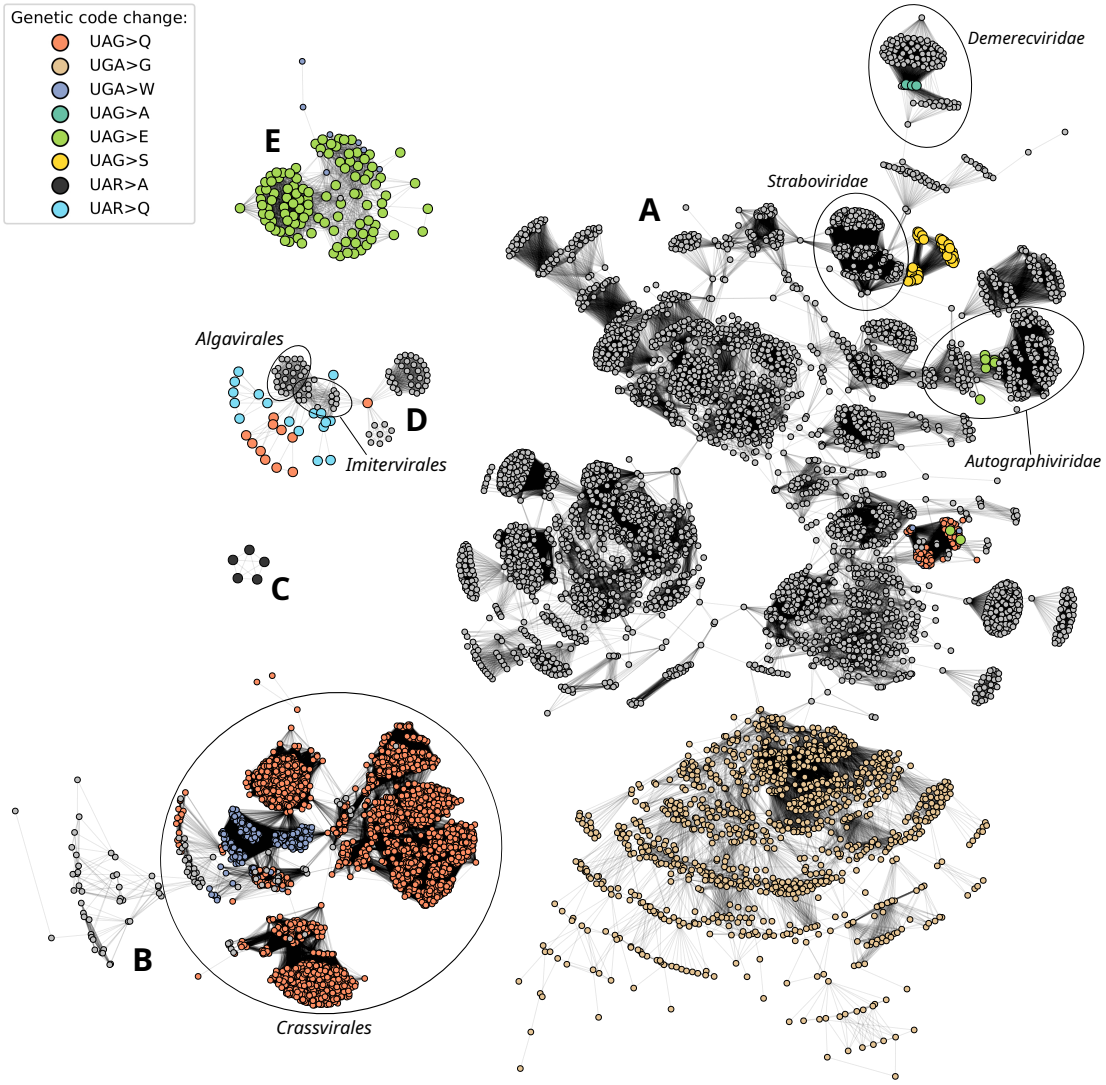


FIGURE 3.1: Gene-sharing network of the 5,585 IMG/VR genomes likely using alternative genetic codes (colored nodes), 4,433 RefSeq *Caudoviricetes* genomes, and 145 RefSeq *Nucleocytoviricota* genomes (grey nodes), generated by vConTACT2. The bigger nodes highlight the groups using alternative genetic codes described in this paper. Selected taxonomic units are marked. Only chosen connected components shown, denoted A to E.

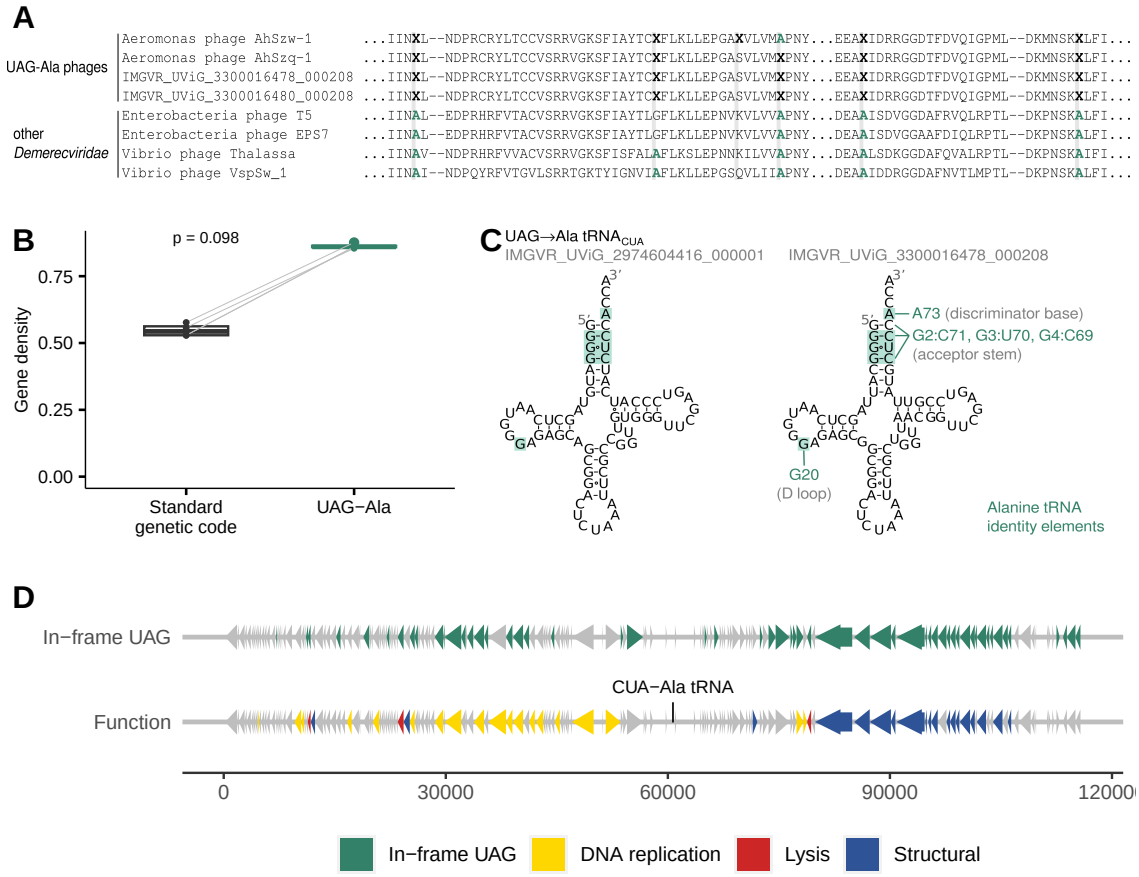


FIGURE 3.2: Reassignment of UAG from stop to alanine in *Shenzhenvirus*, *Demerecviridae*. (A) Multiple sequence alignment of terminase large subunit from the reassigned genomes and 4 closely related *Demerecviridae* genomes. UAGs are represented as X. Selected parts of the alignment containing multiple nearby UAG positions are shown. (B) Gene density under the standard genetic code (black) and under the UAG-Ala code (green). (C) UAG-decoding tRNAs present in the recoded genomes. Alanine tRNA identity elements are highlighted in green [60]. (D) Genomic map of a representative UAG-Ala genome (*Aeromonas* phage AhSzw-1, IMGVR_UViG_2974604416_000001). In the top row, genes containing in-frame UAG codons are highlighted in green; in the bottom row, genes are colored based on their function. The position of the UAG-decoding tRNA is marked.

3.3.2 *Codon reassignments serving as a late-phase switch*

3.3.2.1 *Reassignment of the stop codon UAG to alanine in Shenzhenvirus*

Fifteen viral genomes were inferred to translate UAG, a canonical stop codon, as alanine. Of these, four are part of Component A of the gene-sharing network (Fig. 3.1; the remaining 11 genomes are singletons or do not cluster with any RefSeq genomes). Below, we focus on these four genomes.

Two of the genomes are part of NCBI RefSeq (NC_047949, *Aeromonas* phage AhSzq-1 and NC_047950, *Aeromonas* phage AhSzw-1) and are classified as *Demerecviridae, Shenzhenvirus*. The other two were assembled from metagenomes of freshwater bacterial communities from Singapore. All four genomes are estimated to be 100% complete in IMG/VR and are between 110 and

120 kb long. Consistent with the NCBI annotation, the reassigned genomes cluster with other *Demerecviridae* phages in our gene-sharing network (Fig. 3.1). *Demerecviridae* are a family of lytic phages with a non-enveloped, head-tail structure, with bacteriophage T5 being a typical representative [145].

Fig. 3.2A shows an example multiple sequence alignment of the terminase large subunit from the recoded genomes and four closely related *Demerecviridae* genomes. The alignment shows that UAG codons tend to occur at positions conserved for alanine and other small nonpolar amino acids. The gene density increases from an average of 54.8% under the standard genetic code to 86.2% when considering a genetic code with the UAG-Ala reassignment (Fig. 3.2B), although the difference is not statistically significant due to the low number of genomes ($p = 0.098$). All four genomes encode a CUA-anticodon tRNA with sequence elements consistent with alanine identity, including the major determinant of alanine identity, the $G_3 \cdot U_{70}$ wobble pair [146], [147] (Fig. 3.2C). The genes containing in-frame UAG codons are highly clustered in all four genomes ($p \leq 0.0002$ after Benjamini-Hochberg correction for multiple testing). As a representative, Fig. 3.2D depicts the genome of *Aeromonas* phage AhSzw-1. We annotated the genes in this genome with functional predictions by running HMMER against the VOG database of viral protein profiles. The UAG-containing genes are significantly enriched in structural genes ($p = 3.9 \cdot 10^{-4}$). All three lysis-related genes also contain in-frame UAG codons, although the enrichment is not statistically significant ($p = 0.185$).

In summary, the presence of the UAG-decoding tRNAs with alanine identity elements, along with the association of in-frame UAG codons with structural and lysis genes, strongly suggests that these four *Demerecviridae* bacteriophages utilize the change from the standard to the UAG-Ala genetic code as a late-phase switch.

3.3.2.2 *Reassignment of the stop codon UAG to serine in a group of Straboviridae phages*

43 genomes were inferred to decode the canonical stop codon UAG as serine. In the gene-sharing network, all of them are clustered in Component A, connected to the *Straboviridae* family (Fig. 3.1), encompassing the "T4-like" phages [148]. The UAG-Ser genomes seem to be most closely related to genus *Schizotequatrovirus*, as three out of the seven RefSeq genomes connected to them in the gene-sharing network belong to this genus (the remaining four are classified as *Ishigurovirus*, *Krischvirus*, *Cinqassovirus*, and one is classified only as *Straboviridae*). All 43 genomes originate from metagenomic samples from human oral cavity. Based on a CRISPR spacer match, IMG/VR predicts the host to be *Pasteurellaceae* bacteria for 36 of the 43 genomes, and more specifically, *Haemophilus* for 18 of the 36. The contigs are between 4 and 150 kb long and between 2.56 and 100% complete; the estimated length of the complete genomes ranges from 150 to 225 kb.

Fig. 3.3A shows an example multiple sequence alignment of the baseplate wedge protein gp6 from three of the reassigned genomes, three closely related *Straboviridae* genomes connected to the recoded genomes in the gene-sharing network, and two more distantly related *Straboviridae* genomes, showing that the in-frame UAGs tend to appear at positions conserved for serine and other small non-polar amino acids. The gene density increases from an average of 61.0% under the standard genetic code to 93.4% when considering a UAG-Ser code (Fig. 3.3B; $p < 2.2 \cdot 10^{-16}$). 21 of the 43 genomes encode a CUA-anticodon tRNA with sequence features consistent with serine identity, including the long variable loop (Fig. 3.3C) [60], [149]. We also identified a homologue of

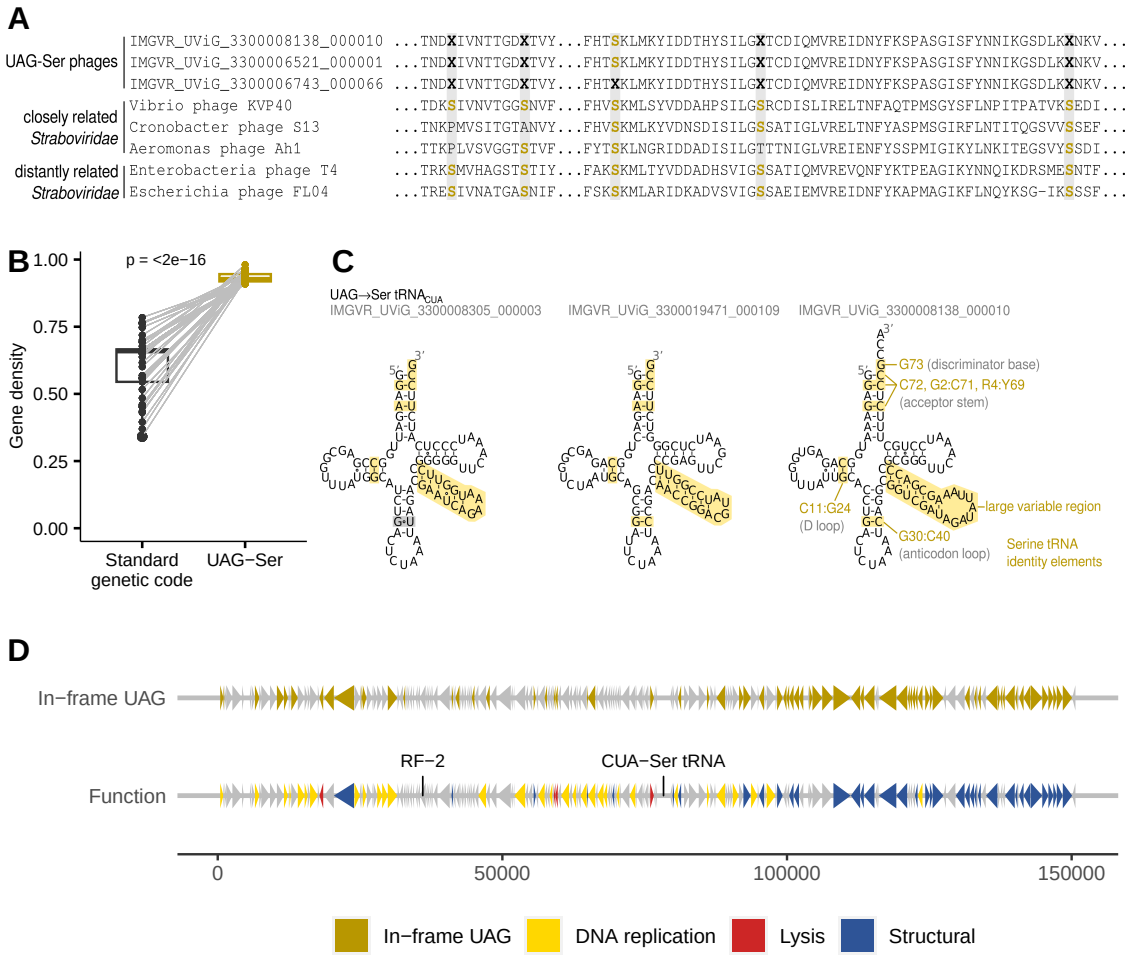


FIGURE 3.3: Reassignment of UAG from stop to serine in a group of *Straboviridae*. (A) Multiple sequence alignment of baseplate wedge protein gp6 from three recoded genomes, three closely related *Straboviridae* genomes (connected to the recoded genomes in the gene-sharing network), and two more distantly related *Straboviridae* genomes. UAGs are represented as X. Selected parts of the alignment containing multiple nearby UAG positions are shown. (B) Gene density under the standard genetic code (black) and under the UAG-Ser code (ochre). (C) UAG-decoding tRNAs present in the recoded genomes. Serine tRNA identity elements are highlighted in yellow [60]. Sequence features not matching serine identity elements are highlighted in grey. (D) Genomic map of a representative UAG-Ser genome (IMGVR_UViG_3300008138_000010, estimated completeness 100%). In the top row, genes containing in-frame UAG codons are highlighted in ochre; in the bottom row, DNA-replication, structural, and lysis genes are colored yellow, blue, and red, respectively. The positions of the UAG-decoding tRNA and a gene encoding release factor 2 are marked.

bacterial release factor 2 (RF-2), which terminates translation at the UGA and UAA codons, in 23 out of the 43 genomes.

Genes containing in-frame UAG codons are significantly clustered ($p \leq 0.05$ after Benjamini-Hochberg correction for multiple testing) in 62.5% of the genomes (20/32; genomes with less than 10 switches in UAG usage were excluded). As an example, Fig. 3.3D shows the genome IMGVR_UViG_3300008138_000010. The UAG-containing genes are highly clustered in this genome ($p = 10^{-4}$). Functionally, these genes are significantly enriched for structural proteins ($p = 2.31 \cdot 10^{-8}$) and depleted in functions related to DNA replication ($p = 0.005$). All four lysis-related genes also contain in-frame UAG codons, although the enrichment is not statistically significant ($p = 0.146$).

To summarize, the consistent alignment of the UAG codons with serine positions, the significant increase in gene density, the presence of UAG-decoding tRNAs with serine identity elements, and the association between UAG usage and structural genes all suggest that these *Straboviridae* phages utilize a UAG-Ser genetic code as a late-phase switch. However, why do they also encode RF-2? It has been hypothesized that UAG-recoded phages infect UGA-recoded hosts [136], which do not encode their own RF-2. The phage-encoded RF-2 could then disrupt host translation, preventing the effective production of host-encoded RF-1, which terminates translation at the UAA and UAG codons, and thereby facilitating an effective UAG-recoding. Indeed, in the UAG-Ser phages, the UGA stop codon is used much less frequently than the other two stop codons (mean proportion 11.8%, for genes predicted using the standard genetic code). However, there are no known UGA-recoded *Pasteurellaceae* species [59], and thus the exact function of the RF-2 present in the UAG-Ser genomes remains unknown.

3.3.2.3 *Reassignment of the stop codon UAG to glutamate in two distinct Caudoviricetes clades*

127 genomes were inferred to translate the canonical stop codon UAG as glutamate. These genomes fall into three distinct clusters in the gene-sharing network (Fig. 3.1). A majority, 106 of the 127 genomes, form the bulk of Component E and likely use the UAG-Glu genetic code throughout their genomes; we will discuss these later together with the other genome-wide genetic codon reassignments. Six genomes cluster with the *Studiervirinae* subfamily, *Autographiviridae* family in Component A of the gene-sharing network. Additionally, two genomes cluster with the unclassified *Bacillota*-infecting phages in Component A that use the UAG-Gln and UGA-Trp genetic codes (Fig. 3.1). The remaining 13 genomes are not members of any of the components shown in Fig. 3.1.

We will first focus on the two genomes that cluster with the unclassified *Bacillota*-infecting UAG-Gln and UGA-Trp phages (Fig. 3.4). Both genomes come from sheep rumen microbial communities, are about 170 kb long and 98.36-100% complete. As shown in Fig. 3.4A, the UAG codons often, though not always, align with positions where glutamate is conserved. They also appear at different positions than the reassigned codons of the UAG-Gln and UGA-Trp genomes within the same gene-sharing cluster. Gene density increases from an average of 59.6% under the standard genetic code to an average of 87.3% under a genetic code where the UAG codon specifies an amino acid (Fig. 3.4B). Both genomes contain a CUA-anticodon tRNA, predicted as Glu isotype by tRNAscan-SE 2.0 [152]. These tRNAs possess some of the glutamate identity elements: A37, the base pair U11-A24, the base-triple U13-G22:A46, the absence of residue 47 [150], and A73 (which

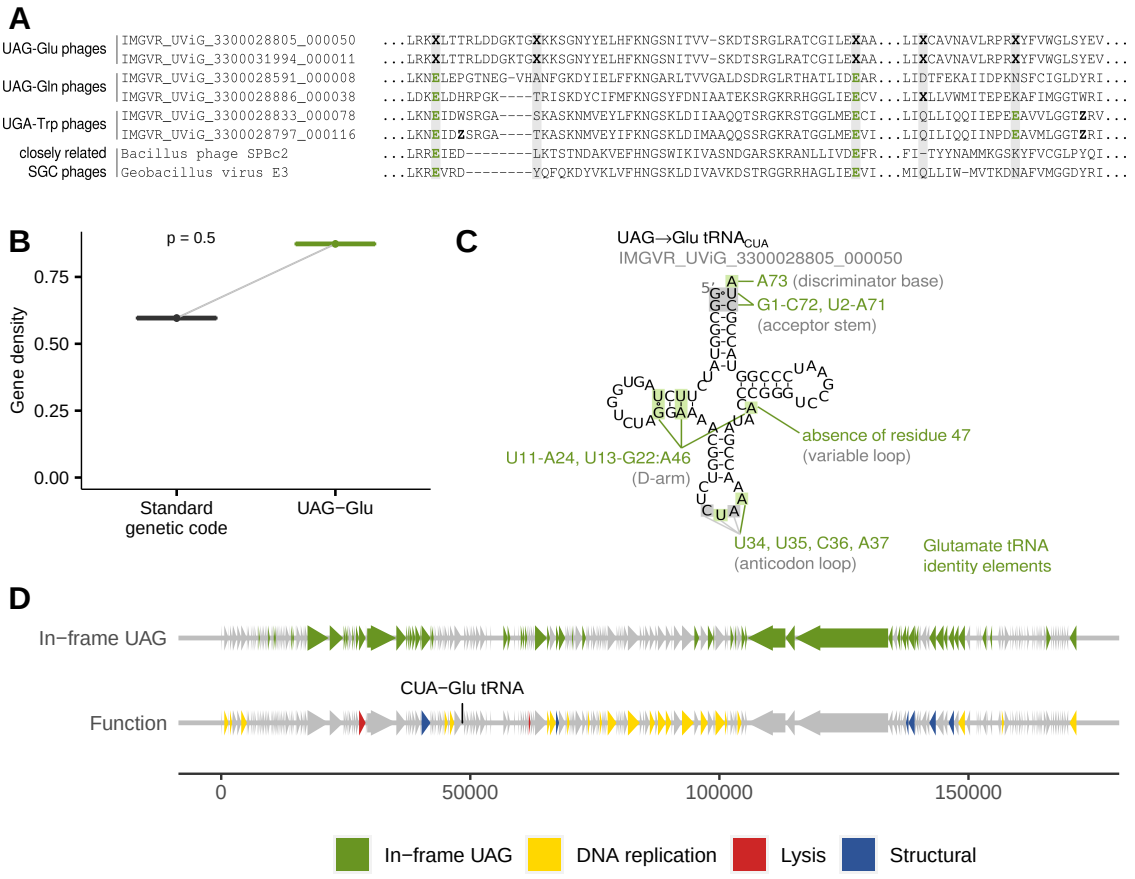


FIGURE 3.4: Reassignment of UAG from stop to glutamate in two unclassified *Bacillota*-infecting phages. Phages in the same cluster also use the UAG-Gln and UGA-Trp genetic codes (see Fig. 3.1). (A) Multiple sequence alignment of terminase large subunit from the UAG-Glu genomes, two closely related UAG-Gln genomes, two closely related UGA-Trp genomes, and two closely related genomes using the standard genetic code (SGC). UAGs are represented as X, UGAs as Z. Selected parts of the alignment containing multiple nearby UAG and UGA positions are shown. (B) Gene density under the standard genetic code (black) and under the UAG-Glu code (green). (C) UAG-decoding tRNA present in both recoded genomes. Glutamate tRNA identity elements are highlighted in green if present, in grey if not present [60], [150], [151]. (D) Genomic map of a representative genome (IMGVR_UViG_3300028805_000050, estimated completeness 98.36%). In the top row, genes containing in-frame UAG codons are highlighted in green; in the bottom row, genes are colored based on their predicted function. The position of the UAG-decoding tRNA is marked.

improves recognition by glutamyl-tRNA-synthetase in the presence of the CUA anticodon [151]). However, they lack others, such as the G1-C72 and U2-A71 base pairs in the acceptor stem [150] (Fig. 3.4C). These tRNAs differ from those of the UAG-Gln and UGA-Trp phages in the same cluster (Supp. Fig. S3.3). The tRNAs have only a 4bp anticodon stem, however, functional tRNAs with a 4bp anticodon stem have been observed previously [153]. In both genomes, the UAG-using genes are highly clustered ($p = 10^{-4}$ in both cases; Fig. 3.4D). Functionally, these genes are enriched in structural properties ($p = 0.0029$) and depleted in DNA synthesis-related functions ($p = 0.0045$) (Fig. 3.4D).

The six *Studiervirinae* phages mostly originate from human large intestine metagenomic samples and are around 43 kb long. Fig. 3.5A presents an example multiple sequence alignment of the tail tubular protein gp12 from two of the recoded genomes, two closely related *Studiervirinae* genomes, and two more distantly related *Autographiviridae* genomes. The alignment shows that UAG codons tend to appear in positions where closely related viruses use glutamate or similar amino acids, such as aspartate or glutamine. The gene density increases from an average of 60.0% under the standard genetic code to 92.1% under a UAG-rewired code ($p = 0.036$, Fig. 3.5B). Five out of the six genomes encode a CUA-anticodon tRNA. Similar to the two genomes discussed above, these tRNAs have some, though not all, glutamate identity elements (Fig. 3.5C). In five out of the six genomes, the UAG-using genes are significantly clustered ($p < 0.05$; the one genome without significant clustering contains only 12 genes). In the example genome, IMGVR_UViG_3300045988_015791, the UAG-using genes are significantly enriched in structural genes ($p = 0.0015$) and significantly depleted in genes related to DNA replication ($p = 0.023$) (Fig. 3.5D).

In summary, in both clades, the significant increase in gene density, the presence of a UAG-decoding tRNA, and the alignment of UAG-containing genes with structural genes suggest that these genomes use a UAG-reassigned genetic code as a late-phase switch. Our Codetta analysis indicates that the reassignment is likely UAG to glutamate. However, this conclusion is tentative due to the absence of some glutamate identity elements in the corresponding tRNAs and the limited number of genomes where this reassignment was observed.

3.3.3 *Genome-wide genetic code changes*

3.3.3.1 *Reassignment of the stop codon UAG to glutamate in a group of unclassified bacteriophages*

The majority of the genomes predicted to translate UAG as glutamate are clustered in Component E of the gene-sharing network, along with two RefSeq genomes infecting aquatic Gram-negative bacteria: NC_004735 Rhodothermus phage RM378 (classified only as *Caudoviricetes*) and NC_055915 Flavobacterium phage vB_FspM_immuto_2-6A (the only RefSeq member of genus *Immutovirus*). Consistently, most of the 106 IMG/VR genomes in Component E come from metagenomic samples from wastewater or landfills. The contigs are between 5 and 262 kb long and between 2.3 and 86.4% complete. The estimated length of the complete genome ranges from 70 to 651 kb, with the mean of 200 kb.

Gene density increases from an average of 30.5% under the standard genetic code to 92.6% under the alternative code ($p < 2 \cdot 10^{-16}$; Fig. 3.6A). Ten out of the 106 genomes encode at least one CUA-anticodon tRNA. These tRNAs contain some sequence features consistent with glutamate

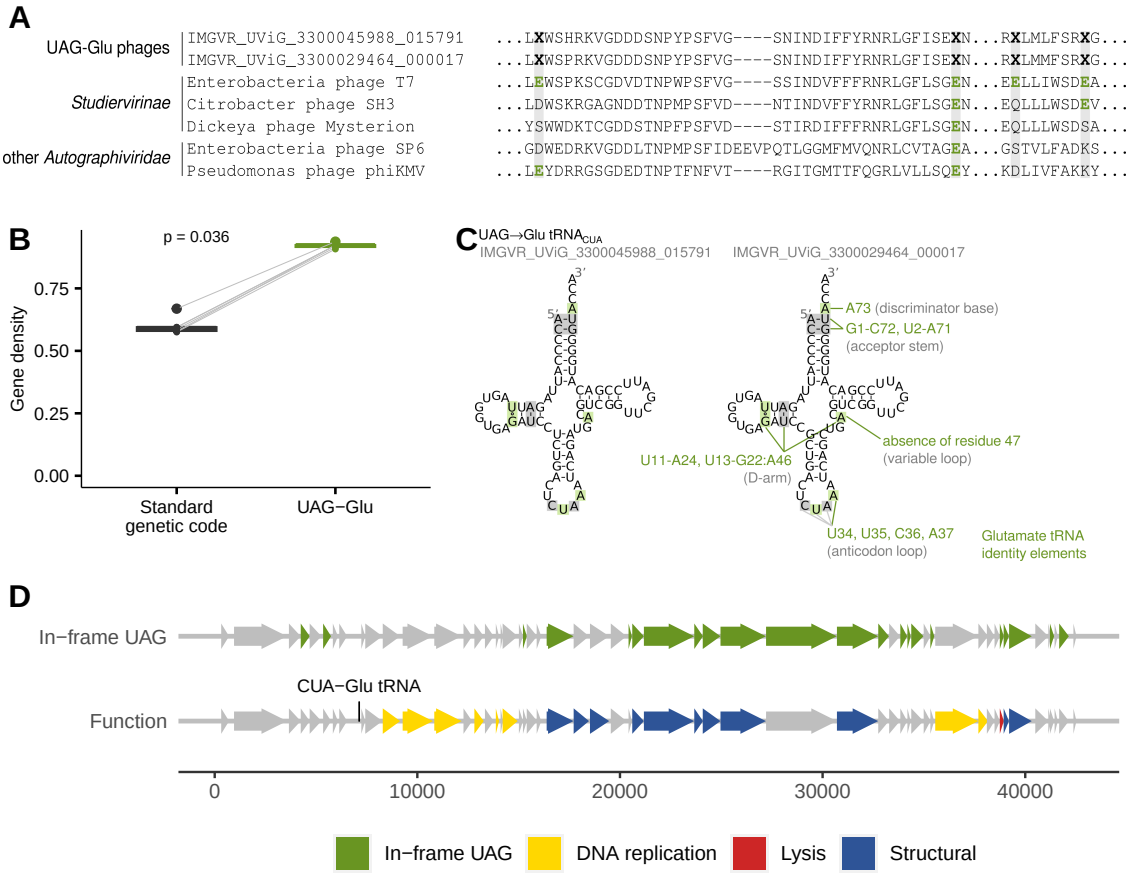


FIGURE 3.5: Reassignment of UAG from stop to glutamate in members of *Studiervirinae*, *Autographiviridae*. (A) Multiple sequence alignment of tail tubular protein gp12 from two UAG-Glu genomes, two closely related *Studiervirinae* genomes, and two more distantly related *Autographiviridae* genomes. UAGs are represented as X. Selected parts of the alignment containing multiple nearby UAG positions are shown. (B) Gene density under the standard genetic code (black) and under the UAG-Glu code (green). (C) UAG-decoding tRNAs present in the recoded genomes. Glutamate tRNA identity elements are highlighted in green if present and in grey if not present [60], [150], [151]. (D) Genomic map of a representative genome (IMGVR_UViG_3300045988_015791, estimated completeness 100%). In the top row, genes containing in-frame UAG codons are highlighted in green; in the bottom row, genes are colored based on their predicted function. The position of the UAG-decoding tRNA is marked.

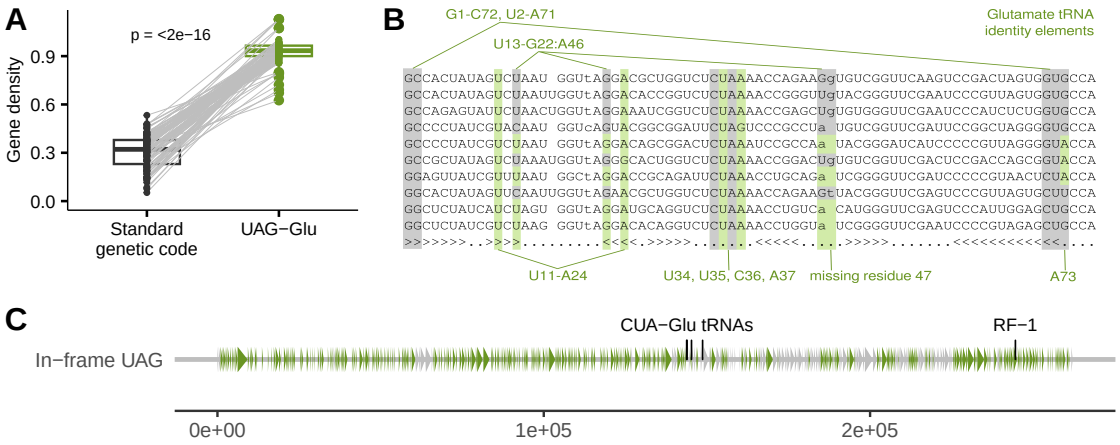


FIGURE 3.6: Reassignment of UAG from stop to glutamate in a group of unclassified phages (Component E of the gene-sharing network). (A) Gene density under the standard genetic code (black) and under the UAG-Glu code (green). (B) Multiple-sequence alignment of unique sequences of the UAG-decoding tRNAs present in the recoded genomes. Glutamate tRNA identity elements [60], [150], [151] are highlighted in green if present and in grey if not. The last row denotes the secondary structure in the dot-bracket notation. (C) Genomic map of a representative genome (IMGVR_UViG_3300020048_000004; estimated completeness 86.41%). Genes containing in-frame UAG codons are highlighted in green. The positions of the three UAG-decoding tRNAs and the gene encoding RF-1 are marked.

identity but not all (Fig. 3.6B). Three genomes possess a homolog of release factor 1 (RF-1), which terminates translation at stop codons UAA and UAG, and 11 genomes encode an aminoacyl tRNA synthetase class Ib, which includes the glutamyl- and glutaminyl-tRNA synthetases [154]. Unlike the genetic code changes described so far, this alternative genetic code appears to be used throughout the genome: the mean proportion of genes containing at least one in-frame UAG codon is 86.1% (ranging from 33.3% to 100%) (Fig. 3.6C). The UAG-using genes are significantly clustered in only 29.1% of the genomes, and they are not significantly enriched or depleted for any specific function. In particular, in-frame UAG codons are prevalent both in typically late-phase genes (structural and lysis-related) and in early-phase genes, such as nucleotide metabolism and DNA replication genes. For instance, in genome IMGVR_UViG_3300020048_000004, 20 out of 25 DNA replication-related genes contain in-frame UAG codons. It is thus unlikely that the recoding appears to be genome-wide only because of the incompleteness of the genomes.

In summary, the substantial increase in gene density together with the presence of UAG-decoding tRNAs strongly suggest that the canonical stop codon UAG encodes an amino acid in these genomes. Codetta consistently predicts it is translated as glutamate, which is further supported by the presence of aminoacyl tRNA synthetase class Ib in some genomes. However, the sequences of the corresponding CUA-anticodon tRNAs are not completely consistent with glutamate identity. The presence of RF-1, responsible for terminating translation at the UAG and UAA codons, is puzzling, as it would likely disable, or at least greatly diminish the effectiveness of the UAG-recoding. In all three genomes where RF-1 is present, its sequence contains numerous in-frame UAG codons, suggesting that its production may be self-regulated. Alternatively, it may have mutated to no longer effectively recognize the UAG stop codon. Further investigation is needed to elucidate the precise mechanisms underlying this genetic code reassignment.

It also remains unclear whether the codon reassignment is specific to the viruses or represents an adaptation to the host's genetic code. Unfortunately, we do not know the host organism of the bacteriophages in Component E. The absence of known examples of UAG-recoded bacteria or archaea [59], though, decreases the probability that this codon reassignment evolved as an adaptation to the host's genetic code. On the other hand, if the genetic code change is host-independent, it would be the first known instance of a virus using a genetic code different from its host throughout the genome. How such a virus could operate, given its reliance on the host translation machinery, especially early in infection, is unclear.

3.3.3.2 *Reassignment of the stop codons UAA and UAG to alanine in a group of unclassified bacteriophages*

Five genomes were predicted by Codetta to translate the stop codon UAA as alanine. These genomes are the sole members of Component C (Fig. 3.1), making their taxonomy unclear; in IMG/VR they are classified only as *Caudoviricetes*. The contigs are between 29 and 45.5 kb long and their estimated completeness ranges from 14 to 22%. All five genomes originate from wastewater-treating anaerobic digesters.

While only the UAA-Ala reassignment met Codetta's parameters for a reliable prediction, there is good evidence that UAG is also translated as alanine. The probability of UAG being decoded as alanine is 99.3% in four of the five genomes and 83.0% in the remaining genome (our threshold for a reliable prediction was 99.99999%; Methods). Additionally, in-frame UAGs frequently occur within genes alongside UAAs (Fig. 3.7A).

To further test the hypothesis that the UAG codon is also reassigned to alanine, we aimed to compare gene density under the UAA-Ala code with the UAR-Ala code (where R denotes A or G). Predicting genes using the UAA-Ala code is unfortunately not possible with Prodigal, the gene prediction tool we used [155], as it does not allow the specification of a custom genetic code and none of the NCBI genetic codes reassigns only the UAA stop codon. Instead, we compared gene density under the UAR-Ala code with a URA-Ala code. In both cases, we assumed that two of the canonical stop codons encode alanine, but while there is no evidence for a UGA reassignment, we suspect the UAG reassignment might be real. As expected, we observe a significant increase in the gene density between the URA-Ala and UAR-Ala codes (Fig. 3.7B).

Additionally, all five genomes encode a CUA-anticodon tRNA with alanine identity elements, along with two to three UUA-anticodon tRNAs, also consistent with alanine identity (Fig. 3.7C). Four out of the five genomes also encode a homolog of RF-2. The genes that use in-frame UAA and UAG codons are not significantly clustered in any of the 5 genomes ($p > 0.084$ in all of them) and in-frame UAR codons are used by between 75.7 and 82.4% of genes (Fig. 3.7D). The observed pattern of a genome-wide reassignment does not seem to be an artifact of the low level of completeness, as we have identified multiple typically early-phase genes, such as DNA polymerase or DNA primase, within these genomes that also contain in-frame UAA and UAG codons (Fig. 3.7A). This makes it unlikely that this codon reassignment serves as a late-phase switch.

The UAR to alanine reassignment has not been reported in any organisms or organellar genomes to date. This, combined with the presence of tRNAs that could facilitate the genetic code change and the virus-encoded RF-2, suggests that the reassignment might be virus-specific and

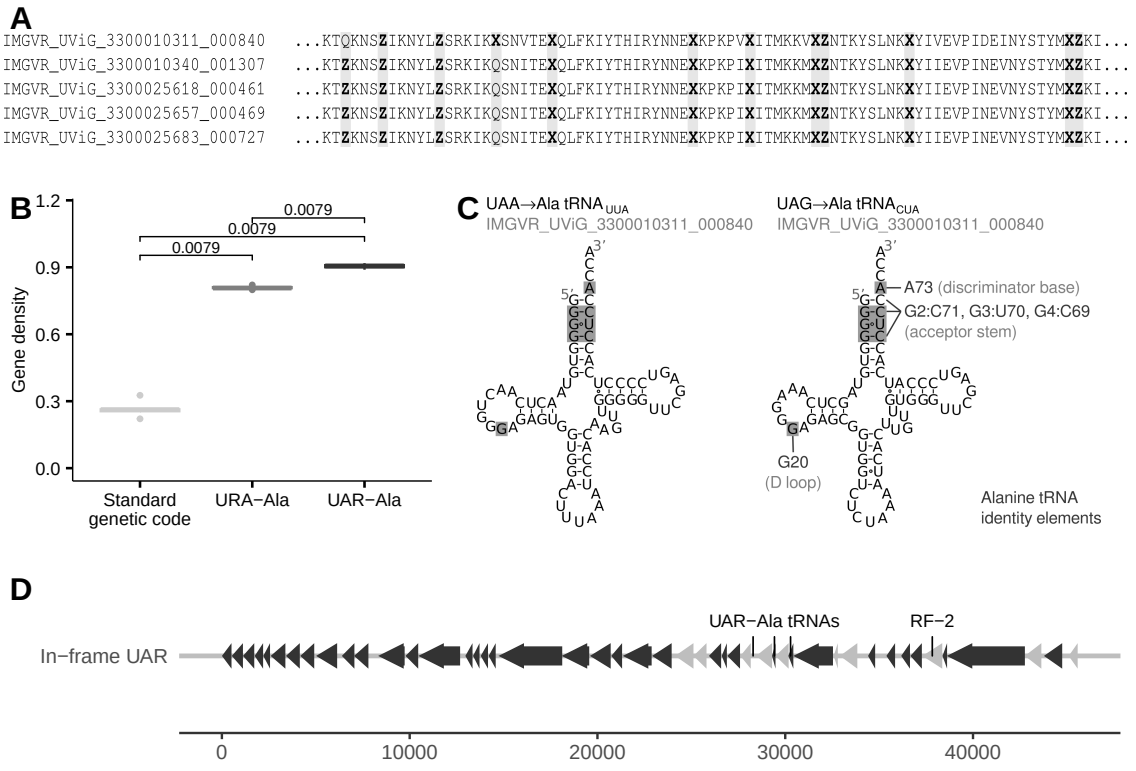


FIGURE 3.7: Reassignment of UAA and UAG from stop to alanine in a group of unclassified bacteriophages. (A) Multiple sequence alignment of DNA polymerase from the recoded genomes. UAA codons are represented as X, UAG codons as Z. Selected part of the alignment containing multiple nearby UAA and UAG positions shown. (B) Gene density under the standard genetic code, under the URA-Ala code, and under the UAR-Ala code. (C) Example UAA- and UAG-decoding tRNAs present in the recoded genomes. The CUA-anticodon tRNA sequence (right) is the same in all five genomes, the UUA-anticodon tRNAs (left) differ slightly in regions not relevant for alanine identity. Alanine tRNA identity elements are highlighted in grey [60]. (D) Genomic map of a representative UAR-Ala genome (IMGVR_UViG_3300010311_000840; estimated completeness 22.1%). Genes containing in-frame UAA or UAG codons are highlighted in black. Positions of the UAA- and UAG-decoding tRNAs and the gene encoding RF-2 are marked.

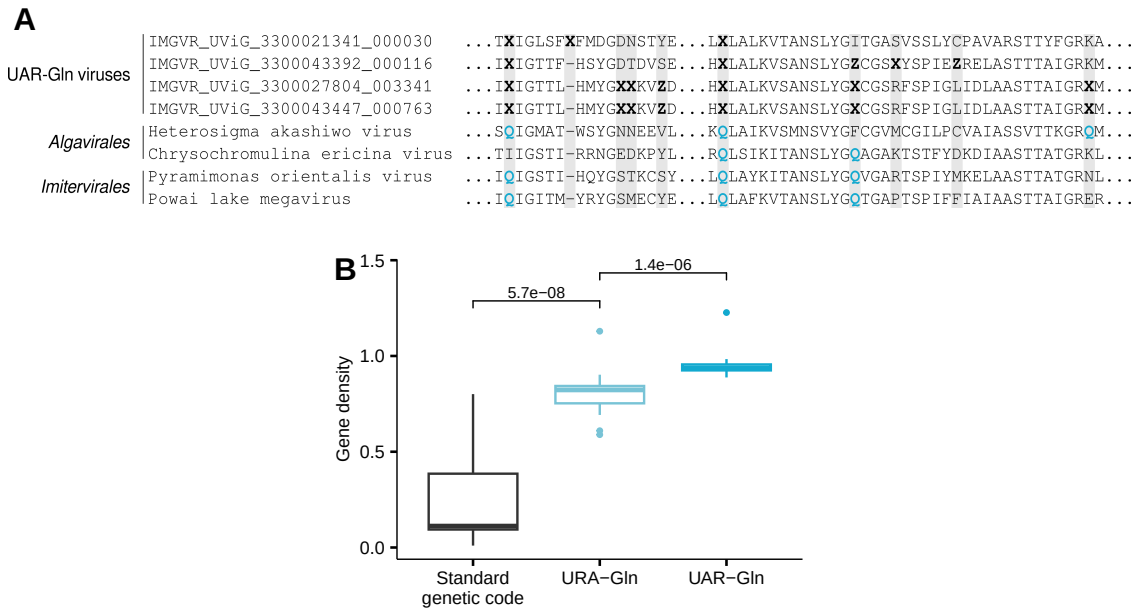


FIGURE 3.8: Reassignment of the UAA and UAG codons from stop to glutamine in a group of *Megaviricetes* viruses. (A) Multiple sequence alignment of DNA polymerase from four of the reassigned genomes, two closely related *Algavirales* genomes, and two closely related *Imitervirales* genomes. UAA codons are represented as X, UAG codons as Z. Selected parts of the alignment containing multiple nearby UAA and UAG positions shown. (B) Gene density under the standard genetic code, under the URA-Gln code, and under the UAR-Gln code.

not host-mediated. Notably, the RF-2 sequences do not contain the reassigned codons, indicating that they could potentially be produced under the host genetic code. Alternatively, these viruses may infect a yet unknown organism which also utilizes the UAR-Ala genetic code.

3.3.4 *Two distinct genetic code changes in Megaviricetes*

All the genetic code changes described so far have been observed in bacteriophages. However, we also detected 67 genomes of large eukaryotic viruses from the class *Megaviricetes* with predicted alternative genetic codes. In all cases, the predicted changes were UAG to glutamine, UAA to glutamine, or UAR to glutamine. 28 of these genomes are clustered in Component D of the gene-sharing network (Fig. 3.1), along with 78 RefSeq *Megaviricetes* genomes. We believe these 28 genomes represent two distinct genetic code changes: UAG-Gln and UAR-Gln.

3.3.4.1 *Reassignment of the UAA and UAG codons to glutamine*

Based on our findings, 17 out of the 28 IMG/VR viruses in Component D utilize a UAR-Gln genetic code. Within this subset, Codetta accurately predicts this coding in three instances, whereas in the remaining 14 cases, it only infers the reassignment of UAA. However, in-frame UAG codons occur within genes alongside UAAs (Fig. 3.8A) and gene density under the UAR-Gln code is significantly higher than under the URA-Gln code (Fig. 3.8B). Thus, we believe that the UAG codon is reassigned to glutamine too. The relatively low number of in-frame UAGs,

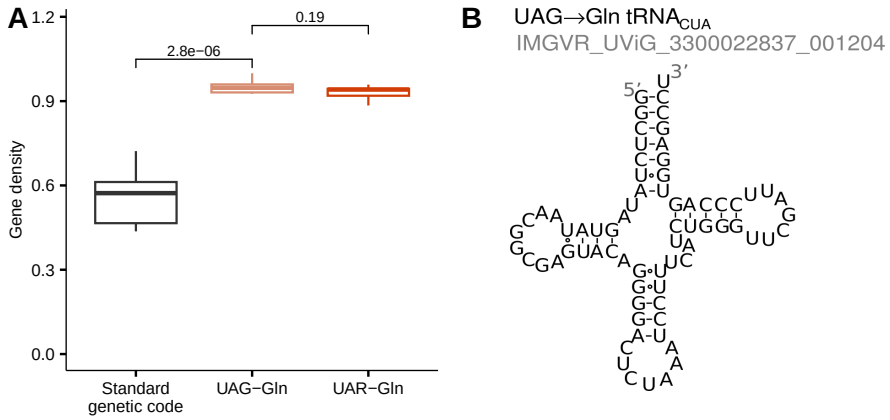


FIGURE 3.9: Reassignment of the UAG stop codon to glutamine in a group of *Megaviricetes* viruses. (A) Gene density under the standard genetic code, the UAG-Gln code, and the UAR-Gln code. (B) UAG-decoding tRNA present in the recoded genomes. No experimentally characterized glutamine identity elements are known for eukaryotes [60] but the tRNA's isotype is predicted as Gln by tRNAscan-SE [152].

compared to UAAs, can be explained by the extremely low GC content of the genomes (mean 19.6%, ranging from 13.3% to 26.4%). There are no tRNAs that could facilitate the genetic code change. However, two of the genomes encode a homolog of eukaryotic release factor 1 (eRF1). The mean proportion of genes using in-frame UAAs or UAGs is 95.4%, which suggests that the genetic code alteration is genome-wide.

All 17 genomes come from freshwater or coastal seawater samples from North America. In the gene-sharing network, they cluster with orders *Algavirales* and *Imitervirales* (Fig. 3.1). Members of the *Algavirales* and *Imitervirales* orders infect eukaryotic algae and protists, respectively. Conveniently, the UAR-Gln genetic code is known to be used by some ulvophycean green algae [156]–[158], various ciliate groups [159]–[164], as well as several other protists [165]–[170]. This, together with the fact that the viruses lack the translational apparatus that could facilitate the genetic code change, suggests that the use of the UAR-Gln genetic code in these viruses is an adaptation to their host.

3.3.4.2 *Reassignment of the UAG stop codon to glutamine*

Based on our evidence, the remaining 11 IMG/VR genomes in Component D use a genetic code in which UAG encodes glutamine, while the meaning of UAA remains unchanged relative to the standard genetic code. We believe this is the case because we do not observe in-frame UAA codons, and gene density does not significantly increase under a code in which both UAG and UAA encode glutamine, compared to a code where UAA is a stop codon (Fig. 3.9A). Six of the 11 genomes encode a CUA-anticodon tRNA, which was assigned glutamine isotype by tRNAscan-SE (Fig. 3.9B). The alternative genetic code appears to be used throughout the genome, as the mean proportion of genes containing in-frame UAG codons is 88.0%.

All of the genomes, except for one, originate from metagenomic samples from saline Ace Lake in Antarctica. In the gene-sharing network, they neighbor *Algavirales* and *Imitervirales* viruses. We hypothesize that this genetic code change might be an adaptation to a host using a UAG-Gln

genetic code. Specifically, we believe these viruses might infect the stramenopile *Bilabrum*, which also uses the UAG-Gln genetic code and whose sequences have been identified in the same metagenomic samples from Ace Lake as these *Megaviricetes* genomes (Marek Eliáš, personal communication).

3.4 DISCUSSION

We performed a computational screen for alternative genetic codes in more than 15.5 million viral genomes and detected eight previously unknown viral clades using alternative genetic codes. Among these eight clades, six different genetic codes are used, with five of them (UAG-Ala, UAG-Ser, UAG-Glu, UAR-Ala, and UAR-Gln) being previously unknown in viruses. In four bacteriophage clades, the alternative genetic code seems to be used as a late-phase switch, in two other bacteriophage clades, the alternative genetic code is used throughout the genome, and the remaining two cases of genetic recoding appear in *Megaviricetes* genomes, likely as an adaptation to a recoded host.

Previously, the phenomenon of dual-coding and genetic code change as a late-phase switch has been described in two groups of large bacteriophages (crAss-like and Lak phages) [136]–[141] and in *Bacillota*-infecting bacteriophages [140]. The reassignments that were suggested to be used by these viruses were UAG-Gln, UGA-Trp, and UGA-Gly. Indeed, in our large-scale screen, we see two groups of UAG-Gln and UGA-Trp recoded phages, corresponding to the crAss-like and the *Bacillota*-infecting phages (Fig. 3.1). Importantly, though, we identified alternative genetic codes that likely serve as a lytic switch in three additional bacteriophage families: *Autographiviridae* (UAG-Glu), *Straboviridae* (UAG-Ser), and *Demerecviridae* (UAG-Ala). Our study thus shows that the ability to utilize a genetic code change as a gene regulation mechanism is widespread among bacteriophages. In addition, we identified several genomes that likely translate the UAG stop codon as glutamate in the same clade of unclassified *Bacillota*-infecting bacteriophages that also contains phages translating UAG as glutamine and UGA as tryptophan. While the evidence for the UAG-Glu reassignment is mixed, if it is confirmed, it would mean the existence of three distinct genetic codes used by this group, suggesting that codon reassignments evolve easily and rapidly in this group of viruses. The reasons for this incredible evolvability of the genetic code in this particular clade of bacteriophages remain unknown and are an exciting direction for a future research.

The molecular details of the genetic code switch also remain unknown. Notably, three out of the four bacteriophage clades that utilize the genetic code change as a late-phase switch do not encode their own release factors, suggesting that host release factors are necessary for translation termination. The host release factors, though, are then expected to compete with phage suppressor tRNAs for binding of the reassigned codons, reducing the efficiency of the recoding. Several possible mechanisms could allow the phages to circumvent this issue. First, the host release factors might be outcompeted due to overexpression of the phage-encoded tRNAs facilitating the genetic code change. Second, it is possible that the phages encode their own equivalents of release factors, which have such low homology to bacterial release factors that they evade detection in sequence comparisons. These phage-encoded release factors could specifically recognize only the stop codons not reassigned in the phages, making the presence of the host release factors

unnecessary. Third, the virus might disrupt the activity of the host release factors in a specific manner, mitigating their interference with the recoding process.

Two additional groups of unclassified bacteriophages, contained in Components C and E of the gene-sharing network (Fig. 3.1), seem to be using an alternative genetic code throughout their genome. In the case of Component E phages, they likely translate the UAG codon as glutamate, in the case of Component C, both UAG and UAA codons seem to be translated as alanine. The hosts of these phages are unknown. While it is possible that these genetic codes are a case of the virus mirroring the genetic code of its host, no known organisms or organellar genomes use these particular genetic code variations. The discovery of such genome-wide recoded phages thus means that either there are two yet unknown prokaryotic groups using alternative genetic codes awaiting discovery, or that these phages, by a mechanism yet unknown, are able to encode most of their genes using a genetic code different from their host.

Finally, we discovered two instances of genetic code change, UAG to glutamine and UAR to glutamine, among large eukaryotic viruses from the *Megaviricetes* class. To our knowledge, they represent the first known instances of genetic recoding among large eukaryotic viruses. As both reassignments are known in protists and algae [156]–[171], which could serve as hosts to these viruses, we believe these genetic code variations might be adaptations of the viruses to their hosts.

As we only investigated in detail the most prevalent codon reassignments proposed by Codetta and those for which there seemed to be the most evidence, it is possible other partially- or completely-recoded genomes are present in the data set. It is also possible Codetta's ability to identify the instances of partial genome recoding would be improved if it was adapted so that it could predict the usage of several distinct genetic codes in different parts of the same genome, as is done by the MgCod algorithm [142], which predicts partial stop codon-recoding based on increase in gene density. Consequently, the actual number of alternative genetic codes utilized by viruses may be even greater than what we have described.

3.5 METHODS

Computational inference of the genetic code with Codetta

For each genome, the genetic code was predicted using Codetta [59] with the following parameters: profile HMM hit value threshold of 10^{-10}, probability threshold to call an amino acid meaning of 0.9999999, and a maximum fraction of observations for a codon coming from a single profile HMM position of 0.01 (`-e 1e-10 -r 0.9999999 -f 0.01`). The protein profile database was built from the multiple sequence alignments in the VOG database, version 218 (https://vogdb.org/) using `hmmbuild --enone`. The `--enone` option turns off effective sequence number determination, resulting in emission probabilities that are closer to the actual amino acid frequencies in the alignments. Multiple sequence alignments with fewer than 50 sequences were excluded, as were profiles VOG00001, VOG00003, VOG00004, VOG00014, VOG00150, VOG00805, and VOG06052, which were observed to cause mispredictions.

Building the gene-sharing network

We downloaded 4,458 complete *Caudoviricetes* and 145 complete *Nucleocytoviricota* genomes from NCBI RefSeq. Using Codetta, we predicted the genetic code for all of them and excluded 25 of the RefSeq *Caudoviricetes* genomes that were predicted to use an alternative genetic code. For the remaining RefSeq genomes and the suspected recoded IMG/VR genomes, we predicted protein-coding genes using Prodigal V2.6.3 [155], assuming the standard genetic code for the RefSeq genomes and the corresponding alternative genetic code for the recoded genomes. These gene predictions were then used to build the gene-sharing network using vConTACT2 0.11.3 [144], using no reference database (`-db None`) and otherwise default parameters. In particular, vConTACT2 clusters the proteins into homologous groups through a two-step process: first, it performs an all-versus-all BLASTP search, and second, it identifies clusters among the significant hits using the MCL algorithm. Genome similarity for each pair of genomes is then defined as the one-tailed p-value of observing at least the observed number of homologous proteins in common by chance.

The network was visualized in the Python igraph library [172], using the Kamada-Kawai layout.

Identification of tRNA and other translational components genes

The tRNA genes were identified using tRNAscan-SE 2.0.12 [152] in the bacterial (`-B`, for the *Caudoviricetes* genomes) or eukaryotic (`-E`, for the *Megaviricetes* genomes) mode.

Release factors were identified by hmmsearch (v3.3.2) using the TIGRFAM [173] release factor models (TIGR00019, TIGR00020, and TIGR03676). Aminoacyl tRNA synthetase genes were identified by hmmsearch using Pfam models PF00579, PF00749, PF00750, PF01406, PF01921, PF09334, PF00133, PF00587, PF01411, PF01409, PF02091, and PF00152.

Gene density estimation

Prodigal V2.6.3 [155], in "single" mode, was used to predict protein-coding genes, assuming either the standard genetic code or the corresponding alternative genetic code (genetic code 15 for UAG-recoded genomes, 6 for UAR-recoded genomes, and 14 for URA-recoded genomes). Gene density was calculated as the total length of all protein-coding genes divided by the length of the genome.

Definition of genes using the reassigned codons

A protein-coding gene was assumed to use the reassigned codons if any of them was located in-frame within the middle 80% of the gene length. This criterion was applied to exclude genes that actually use the standard genetic code but get slightly extended upon the stop codon reassignment.

Clustering of genes using reassigned stop codons

To determine whether the genes using the reassigned codons were significantly clustered in the genome, we first computed the number of "usage switches" – pairs of consecutive genes such that one gene in the pair uses the reassigned codons and the other does not. In other words, we counted the instances where the usage of the reassigned codons switched from "yes" to "no" or vice versa between two neighboring genes. We then generated a null distribution of the number of switches by randomly permuting the order of the genes in the genome 10,000 times and computing the number of switches for each permutation. Genes using the reassigned codons are considered to be significantly clustered in the original genome if the observed number of switches fell within the bottom 5% of the null distribution.

Functional annotation of genes

Genes were functionally annotated by running hmmscan against the VOG database (version 218). In case of several significant hits to the same gene, the profile with the lowest e-value was retained. Gene functions were then manually categorized into "DNA-replication related", "Structural", "Lysis-related", and "Other/Unknown" categories.

Multiple sequence alignments

All multiple sequence alignments were generated using MAFFT v6.864, with the L-INS-i parameters [174], [175].

ACKNOWLEDGMENTS

We thank Yekaterina Shulgina for many useful comments on the project, Shiraz Shah for discussions and his suggestion to use vConTACT2, and Marek Eliáš for sharing information about the *Bilabrum* genetic code. The IMG/VR sequence data were produced by the US Department of Energy Joint Genome Institute (https://www.jgi.doe.gov/) in collaboration with the user community.

3.6 SUPPLEMENTARY MATERIAL

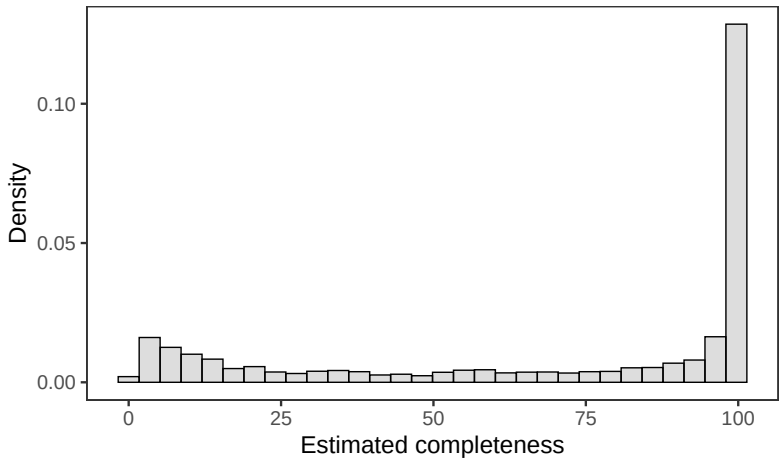


FIGURE S3.1: The estimated completeness (per IMG/VR) of the 5,585 genomes predicted to use alternative genetic codes.

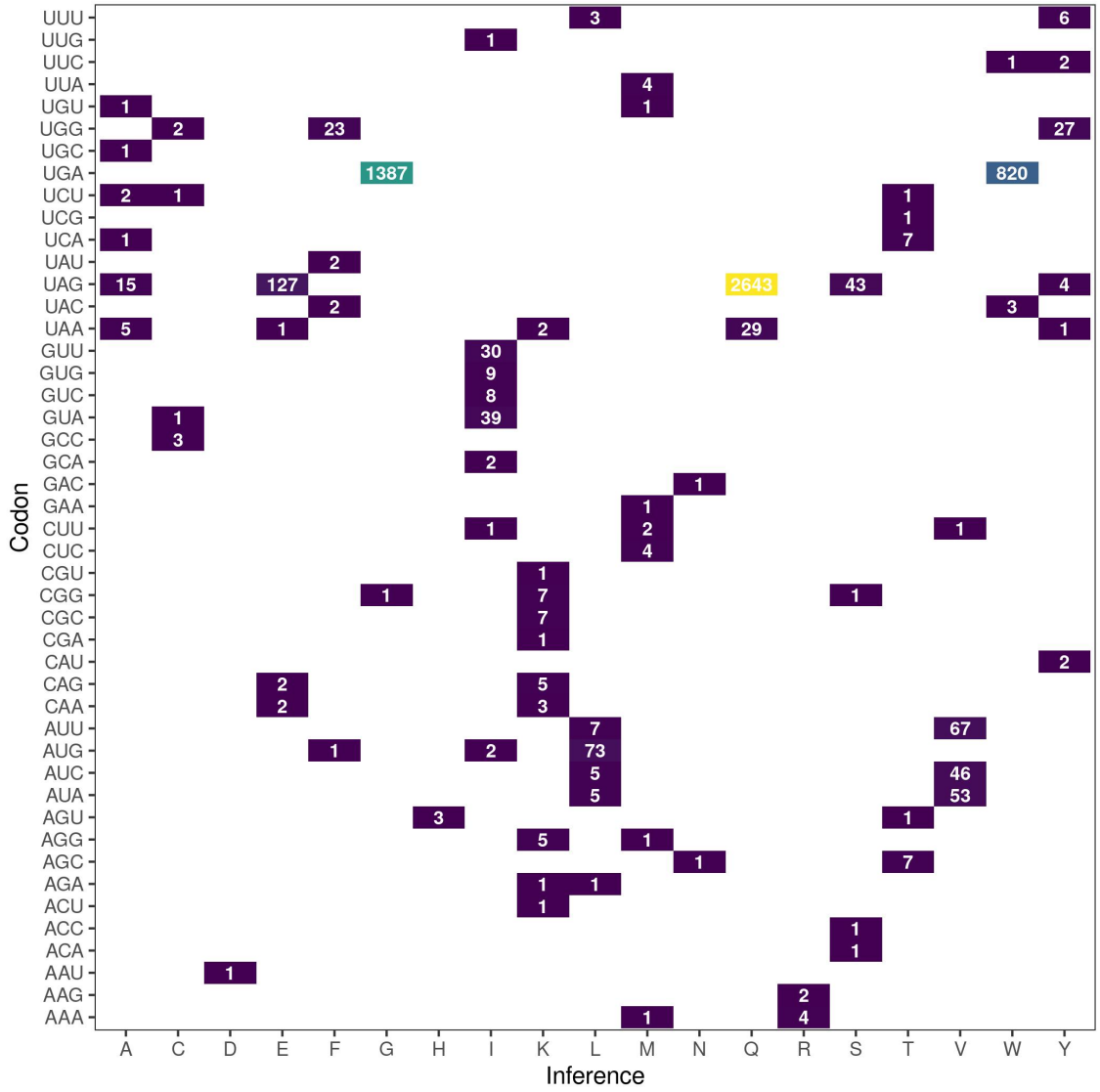


FIGURE S3.2: All candidate codon reassignments predicted in the IMG/VR genomes. Rows denote individual codons, columns the inferred amino acid meaning. Only inferences different from the standard genetic code meaning of the codon are shown. The number within each cell gives the number of occurrences of a given inference, also denoted by the color.

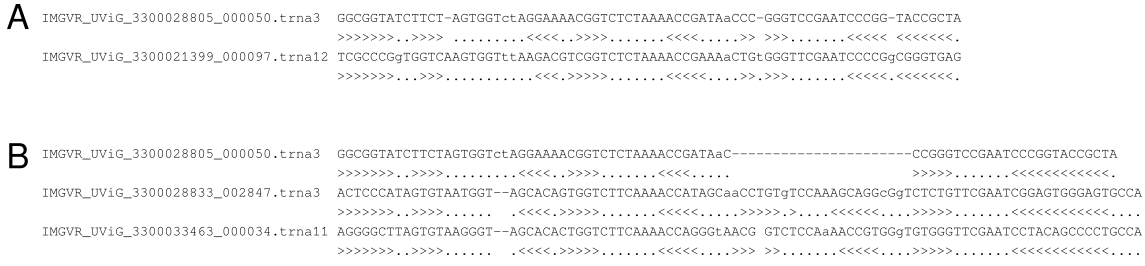


FIGURE S3.3: Alignment of the CUA-anticodon tRNA present in the two unclassified *Bacillota*-infecting UAG-Glu genomes (IMGVR_UViG_3300028805_000050.trna3 and IMGVR_UViG_3300031994_000011.trna3; both have the same sequence) with (A) the most similar CUA-anticodon tRNA from the Component A UAG-Gln phages and (B) the UCA-anticodon tRNAs from the Component A UGA-Trp phages. For each tRNA, the first row specifies the sequence and the second row the secondary structure in the dot-bracket notation.

4

ROBUST GENETIC CODES ENHANCE PROTEIN EVOLVABILITY

Published as: Hana Rozhoňová, Carlos Martí-Gómez, David M. McCandlish & Joshua L. Payne (2024). Robust genetic codes enhance protein evolvability. *PLoS Biology*, 22(5), e3002594. https://doi.org/10.1371/journal.pbio.3002594

Authors' contributions: H.R., C.M.-G., D.M.M. and J.L.P. designed research; H.R. and C.M.-G. performed research; H.R., C.M.-G., D.M.M., and J.L.P. analyzed data; and H.R., C.M.-G., D.M.M. and J.L.P. wrote the paper.

The limits of my language mean the limits of my world.
— Ludwig Wittgenstein

4.1 ABSTRACT

The standard genetic code defines the rules of translation for nearly every life form on Earth. It also determines the amino acid changes accessible via single-nucleotide mutations, thus influencing protein evolvability — the ability of mutation to bring forth adaptive variation in protein function. One of the most striking features of the standard genetic code is its robustness to mutation, yet it remains an open question whether such robustness facilitates or frustrates protein evolvability. To answer this question, we use data from massively-parallel sequence-to-function assays to construct and analyze six empirical adaptive landscapes under hundreds of thousands of rewired genetic codes, including those of codon compression schemes relevant to protein engineering and synthetic biology. We find that robust genetic codes tend to enhance protein evolvability by rendering smooth adaptive landscapes with few peaks, which are readily accessible from throughout sequence space. However, the standard genetic code is rarely exceptional in this regard, because many alternative codes render smoother landscapes than the standard code. By constructing low-dimensional visualizations of these landscapes, which each comprise more than 16 million mRNA sequences, we show that such alternative codes radically alter the topological features of the network of high-fitness genotypes. Whereas the genetic codes that optimize evolvability depend to some extent on the detailed relationship between amino acid sequence and protein function, we also uncover general design principles for engineering non-standard genetic codes for enhanced and diminished evolvability, which may facilitate directed protein evolution experiments and the bio-containment of synthetic organisms, respectively.

4.2 INTRODUCTION

Proteins are the workhorses of the cell. They are the building blocks of cellular infrastructure, they transport molecules, regulate gene expression, and catalyze essential biochemical reactions. How do such protein functions evolve? The classic metaphor of the adaptive landscape is helpful to conceptualize this process [176]. An adaptive landscape is a mapping from genotype space onto fitness or some related quantitative phenotype, which defines the "elevation" of each coordinate in this space. For proteins, genotype space comprises the set of all possible amino acid sequences of a given length [80] and the quantitative phenotypes of these sequences include catalytic activity, folding stability, and binding affinity. The evolution of protein function can then be viewed as a hill-climbing process in such a landscape, in which mutation and natural selection tend to drive evolving populations toward adaptive peaks of improved functionality [177].

Central to this process is evolvability — the ability of mutation to bring forth adaptive phenotypic variation [178], [179]. For short-term, one-step adaptation, evolvability depends on the immediate mutational neighborhood of a protein sequence (Fig. 4.1A). That is, it depends on the amount of adaptive phenotypic variation accessible via point mutation. For longer-term, multi-step adaptation, evolvability depends on the topography of the adaptive landscape. A smooth single-peaked landscape facilitates evolvability, because mutation can easily bring forth

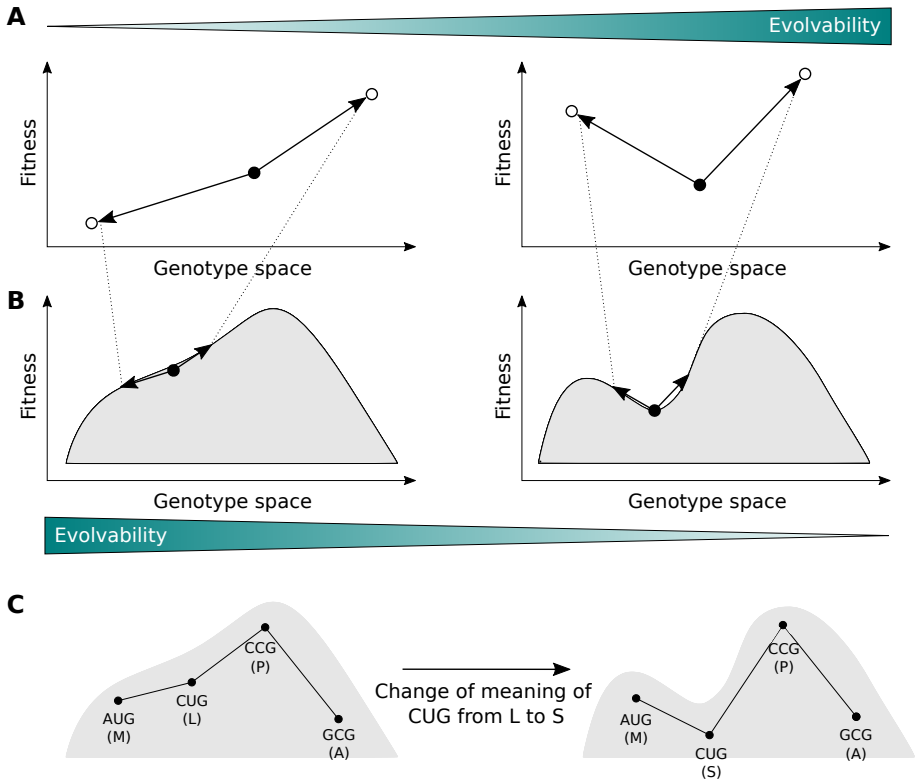


FIGURE 4.1: Evolvability and adaptive landscapes. (A) In one-step adaptation, evolvability depends on the amount of adaptive phenotypic variation accessible via point mutation. Therefore, the genotype shown with a filled circle in the right panel is more evolvable than the one shown in the left panel. (B) Zooming out and considering multi-step adaptation, landscape topography becomes important. Smoother landscapes promote evolvability (left panel), whereas rugged landscapes hinder evolvability (right panel), because an evolving population is more likely to be trapped on a local optimum. (C) Landscape topography is influenced by the genetic code. As a toy model, a sequence consisting of a single codon is shown. Under the standard genetic code, there is a single peak, which is also a global optimum (left panel). If the meaning of the CUG codon is changed from leucine to serine (as is the case in some yeast species [78]), an adaptive valley is formed (right panel). The population now cannot leave the local optimum consisting of the AUG codon without crossing a maladaptive valley.

adaptive phenotypic variation from anywhere in the landscape, except atop the global peak; in contrast, a rugged landscape diminishes evolvability, because its adaptive valleys often preclude the generation of adaptive phenotypic variation [178], [180], [181] (Fig. 4.1B).

What determines whether a protein's adaptive landscape is smooth or rugged? One primary factor is the genetic code an organism uses for translation. Its importance arises because it determines which amino acid changes are accessible via alteration of a single nucleotide. For example, under the standard code, point mutations to the CUG codon can change the amino acid leucine to methionine (AUG), valine (GUG), proline (CCG), glutamine (CAG), and arginine (CGG), but not to any other of the remaining 14 amino acids. A genetic code thus defines which protein sequences are "near" one another in sequence space [81], and which mutational paths to adaptation are closed or open (Fig. 4.1C).

The structure, history, and evolutionary implications of the standard genetic code have fascinated scientists for decades [7], [8], [14], [182], [183]. Given the nearly infinite space of alternatives,

why did life converge on the standard genetic code? What makes it so special? Answers to this question are typically based on comparisons of the properties of the standard genetic code to those of hypothetical, alternative codes [9], [10], of which there are many [96]. Even if one maintains the degeneracy of the standard code, but simply randomizes which amino acids are assigned to which codon blocks, there are $20! \approx 10^{18}$ possible rewirings. By sampling a large number of such rewired codes, one can ask whether a given quantitative property of the standard genetic code has a value higher or lower than expected by chance. For example, using a measure of so-called "error tolerance" based on how well point mutations preserve polar requirement (a measure of hydrophilicity), and taking into consideration mutation bias toward transitions relative to transversions, Freeland and Hurst [10] showed that only one in a million rewired codes preserves the hydrophilicity of amino acids to a greater extent than the standard genetic code. The standard genetic code is thus highly robust to this form of error, in that point mutations and mistranslations tend to cause minor changes to this physicochemical property of amino acids.

What are the implications of code robustness for protein evolvability? By definition, a robust genetic code limits the amount of phenotypic variation that point mutations can cause. However, opinions differ on whether this hinders or facilitates evolvability. Inspired by Fisher's Geometric model [184], early theoretical work argues that code robustness may facilitate protein evolvability exactly because it minimizes the effects of mutations, thus increasing the probability that mutations will be adaptive [185]. Indeed, by analyzing the fitness effects of point mutations to the antibiotic resistance gene TEM-1 β-lactamase and two influenza hemagglutinin inhibitor genes, it has been shown that missense mutations are enriched for adaptive amino acid changes, relative to amino acid changes that require multiple point mutations [13], [186]. In contrast, more recent theoretical work [93], motivated by advances in synthetic biology [40], [187]–[189], argues that protein evolvability can be enhanced by reducing code robustness, because by doing so one can increase the number and diversity of amino acids accessible via point mutation to any codon.

Whether code robustness hinders or facilitates protein evolvability therefore remains an open question. Whereas steps have been taken towards answering this question [13], [94], [95], [186], [190], these studies suffer from at least one of two key limitations. The first is a focus on how missense mutations change the physicochemical properties of amino acids [9], [10], [94], rather than how missense mutations change the phenotype of a protein (e.g., its stability or catalytic activity) or the corresponding fitness of an organism. The second limitation is a lack of suitable data, with studies relying on a purely theoretical model of landscape topography [94], a categorical, rather than quantitative, phenotype [95], an incomplete fitness landscape [13], or assumptions of additivity regarding the combined effects of mutations [190]. We therefore do not know how the structure of a genetic code, standard or otherwise, influences the evolvability of proteins beyond one-step adaptation. This is an important knowledge gap, because protein evolution often proceeds via a sequence of adaptive mutations that improve protein function, as evidenced by comparisons of orthologous sequences [191], [192] and directed protein evolution experiments [193], [194]. Moreover, given the increasing interest in engineering non-standard genetic codes [188], it is desirable to deduce design principles for engineering genetic codes with reduced or enhanced evolvability, as these might be used to form a 'genetic firewall' [195] or accelerate directed evolution [93], respectively.

Here, we overcome the limitations of prior studies using experimental data from massively-parallel sequence-to-function assays [196]. In particular, we use combinatorially complete data, which provide a quantitative characterization of protein phenotype for all possible combinations of 20^L amino acid sequence variants at a small number L of protein sites [197]–[201]. These data facilitate the construction of complete adaptive landscapes without assumptions regarding the combined effects of individual mutations (e.g., additivity). Importantly, the combinatorially-complete nature of these data allows us to construct such landscapes under arbitrary genetic codes. The reason is that, no matter which code we use, we are guaranteed that each of the 4^{3L} possible mRNA sequences can be computationally translated into an amino acid sequence with an experimentally assayed phenotype. We characterize the topographies of six such empirical adaptive landscapes under the standard genetic code, as well as under hundreds of thousands of rewired codes, and perform population-genetic simulations on these landscapes. We show that robust genetic codes tend to produce smooth adaptive landscapes with few peaks and, consequently, allow evolving populations to reach on average higher fitness. Thus, the robustness of a genetic code not only helps to mitigate the potentially deleterious effects of replication and translation errors, but it also transforms the problem of molecular evolution from one that depends on the vicissitudes of individual mutations into one where evolving populations can readily find mutational paths toward adaptation.

4.3 RESULTS

4.3.1 *Data*

We construct empirical adaptive landscapes using six combinatorially-complete data sets for four proteins. The first protein is GB1, a Streptococcal protein that binds immunoglobulin [202], [203]. Wu et al. [197] experimentally assayed the binding affinity of GB1 to immunoglobulin for all $20^4 = 160,000$ amino acid sequences at four protein sites (V39, D40, G41, and V54; Supp. Fig. S4.1), which are known to interact epistatically and influence binding affinity [204]. In particular, they measured the relative frequencies of sequence variants before and after selection for binding immunoglobulin. Binding affinities are then defined as log enrichment ratios (Methods).

The second protein is ParD3, a bacterial antitoxin that is part of the ParD-ParE family of toxin-antitoxin systems, which are commonly found on bacterial plasmids and chromosomes [205]. Such systems comprise a toxin that inhibits cell growth unless bound and inhibited by the cognate antitoxin. Lite et al. [198] experimentally assayed bacterial cell growth for all $20^3 = 8,000$ amino acid sequence variants at 3 sites in ParD3 (D61, K64, E80; Supp. Fig. S4.1), in the presence of its cognate toxin ParE3, as well as a related, but non-cognate toxin ParE2. This resulted in two data sets, one per toxin, in which cell growth was used as a quantitative readout of the degree to which individual ParD3 variants antagonize a given toxin.

The third protein is ParB, a DNA-binding protein crucial for bacterial chromosome segregation [206]. The binding site of ParB, *parS*, is a palindrome of GTTTCAC. Jalal et al. [200] experimentally measured the binding affinity of ParB to the cognate DNA sequence, *parS*, as well as a related DNA-binding site, *NBS* (palindrome of ATTTCCC), for all $20^4 = 160,000$ variants at

four positions (R173, T179, A184, and G201; Supp. Fig. S4.1). This again resulted in two data sets, one per DNA-binding site.

The fourth protein is dihydrofolate reductase (DHFR), an essential metabolic enzyme in *E. coli*. Papkou et al. [201] generated all possible $64^3 = 262,144$ combinations of codons at three positions (A26, D27, L28; Supp. Fig. S4.1) of the corresponding *folA* gene. Missense mutations at these positions are known to confer resistance to the antibiotic trimethoprim [207], [208]. Using a mass-selection experiment, Papkou et al. [201] measured the fitness of each variant in the presence of a sublethal dose of trimethoprim. The majority (89.7%) of the variants are non-functional, in that they are sensitive to trimetophrim.

Following the protein evolution literature [177], [197], [209], we assume that fitness is directly proportional to binding affinity (GB1, ParB) or growth rate (ParD3, DHFR), and will use the term "fitness" generically for all landscapes from now on. Using the raw measurements described above (binding affinities and cell growth), we inferred the fitness values, as well as imputed the missing sequence variants (6.6% of the GB1 data set) using empirical variance component regression [210] (Methods and Supp. Fig. S4.2).

For each of the six data sets, we constructed adaptive landscapes using the standard genetic code, as well as hundreds of thousands of rewired codes. Specifically, we represented each mRNA sequence of length 12 (GB1, ParB-*parS*, ParB-*NBS*) or 9 (ParD-ParE2, ParD-ParE3, DHFR), respectively, as a vertex in a mutational network and connected vertices with an edge if their corresponding sequences differed by a single point mutation [211] (Methods). We labeled each vertex with the fitness of its corresponding translation under a given genetic code, thus defining the "elevation" of each coordinate in genotype space.

4.3.2 *More robust codes cause smoother adaptive landscapes*

How does the robustness of a genetic code influence adaptive landscape topography? To answer this question, we generated 100,000 rewired genetic codes by amino acid permutation, a rewiring scheme that preserves the synonymous codon block structure of the standard genetic code, but randomly permutes the 20 amino acids amongst these blocks [9], [10] (Methods). We quantified the robustness of each code as the proportion of point mutations that do not change the physicochemical properties of amino acids, using the properties defined in ref. [93] (Supp. Fig. S4.3; Methods). According to this measure, the robustness of the standard genetic code is 0.385, meaning that 38.5% of point mutations do not change the physicochemical properties of amino acids. In comparison, the code robustness for the 100,000 rewired codes ranges from 0.257 to 0.462, with a median of 0.336, such that 5.48% of these codes exhibit robustness greater than or equal to the standard code. Therefore, when defining robustness in terms of multiple amino acid properties, the standard genetic code is highly robust, but not surprisingly so [9]. We then constructed an adaptive landscape using each of the 100,000 rewired genetic codes, for each of the six data sets, and characterized the topographies of these landscapes using three measures of landscape ruggedness [212]: the number of adaptive peaks, the prevalence of various types of epistasis, and the proportion of accessible mutational paths to the global peak (Methods). Whereas these three measures are all correlated with one another [213] (Supp. Tab. S4.1), they illustrate different aspects of landscape topography, ranging from local (pairwise epistasis) to

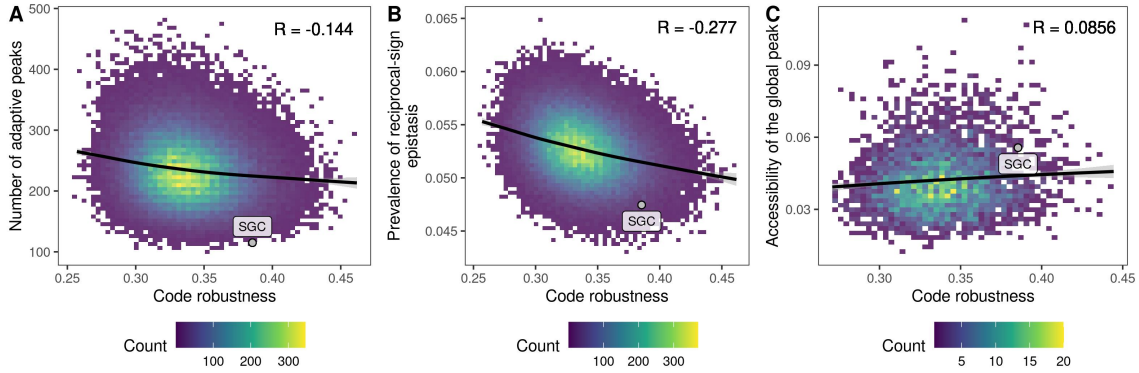


FIGURE 4.2: More robust codes result in smoother adaptive landscapes. Three measures of landscape ruggedness are shown in relation to code robustness, defined as the proportion of point mutations that do not change the physicochemical properties of amino acids. (A) The number of adaptive peaks, (B) the prevalence of reciprocal-sign epistasis, and (C) the proportion of mutational paths to the global peak that are accessible. Panel (C) shows only genetic codes that preserve the size of the global peak relative to the standard genetic code and in which none of the amino acids contained in the global peak (WWLA) are encoded by the split codon block ($n = 3,769$). In each panel, the labeled point denotes the standard genetic code. All results pertain to the GB1 landscape. Analogous results for the ParD, ParB, and DHFR landscapes can be found in Supp. Tab. S4.2.

global (number and accessibility of adaptive peaks). Below, we focus our analyses on the GB1 data, and report analogous results for the ParD, ParB, and DHFR data in Supp. Tab. S4.2.

Adaptive peaks

The number of adaptive peaks is a straightforward measure of landscape ruggedness, and thus of evolvability. The more local peaks a landscape has, the more likely an evolving population is to become trapped on one of these peaks, thus precluding the generation of further adaptive phenotypic variation. Under the standard genetic code, the GB1 landscape comprises 115 adaptive peaks, whereas under the 100,000 rewired codes, the number of adaptive peaks ranges from 97 to 478, with a median of 231.

Fig. 4.2A shows the number of adaptive peaks in relation to code robustness, revealing that more robust genetic codes tend to produce adaptive landscapes with fewer peaks than less robust codes (Pearson's correlation $R = -0.144$, $p < 2.2 \cdot 10^{-16}$). However, the trend is relatively weak, such that for any level of code robustness, there is considerable variation in the number of peaks. For example, for the 5.48% of codes with robustness greater than or equal to the standard code, the number of peaks ranges from 115 to 403. Strikingly, among all 100,000 codes, only 0.037% of the corresponding landscapes have less than or equal the number of peaks in the landscape produced by the standard code. The GB1 landscape is therefore exceptionally smooth under the standard genetic code. To a lesser extent, this is also true for the ParB-*parS* (2.3%) and ParB-*NBS* (4.6%) landscapes. However, this is not true for the ParD-ParE2, ParD-ParE3, and DHFR landscapes, where the number of peaks in the landscape under the standard genetic code lies in the 32.6%, 57.0%, and 20.5% quantile, respectively (Supp. Tab. S4.3). Whether the standard genetic code is exceptional with regard to producing smooth fitness landscapes therefore appears to be data set-specific, a topic we return to later.

Epistasis

Epistasis, where a mutation's effect depends on the genetic background in which it occurs, is a cause of landscape ruggedness [214], [215]. It can be classified into three types — magnitude, simple sign, and reciprocal sign [214] (Methods). Reciprocal sign epistasis occurs when two mutations each have a positive (negative) effect on fitness, but each mutation has negative (positive) effect when introduced in the background of the other mutation. That is, the sign of each mutation's effect flips when introduced in the presence of the other mutation. Reciprocal sign epistasis forms local valleys in an adaptive landscape, which preclude the generation of at least some adaptive phenotypic variation, thus decreasing evolvability. To measure the prevalence of these three types of pairwise epistasis, we randomly sample a large number of squares in each adaptive landscape's underlying mutational network, each of which contains an mRNA sequence variant, two of its single-mutant neighbors, and a double mutant that can be constructed from the single mutants. Based on the fitness values of these four sequences, we classify the type of epistasis the square exhibits (Methods).

Because more robust codes tend to produce adaptive landscapes with fewer adaptive peaks (Fig. 4.2A), we expect landscapes produced under more robust codes to exhibit less reciprocal sign epistasis than landscapes produced under less robust codes. Fig. 4.2B confirms this expectation, showing a negative correlation between reciprocal sign epistasis and code robustness ($R = -0.277$, $p < 2.2 \cdot 10^{-16}$). Similarly, simple sign epistasis, which contributes to landscape ruggedness to a lesser extent than reciprocal sign epistasis, because it involves only a single sign flip, also exhibits a negative correlation with code robustness (Supp. Tab. S4.2). Robust genetic codes therefore diminish the kinds of epistatic interactions that cause landscape ruggedness.

Global peak accessibility

One consequence of landscape ruggedness is that the global adaptive peak may be less accessible to an evolving population, which may instead follow mutational paths to local adaptive peaks. We therefore expect that the global adaptive peaks of landscapes produced under more robust codes will be more accessible than those of landscapes produced under less robust codes. To test this, we quantified the mutational accessibility of the global peak of each landscape by calculating the probability that a randomly chosen, direct mutational path that starts at a randomly chosen mRNA sequence and ends at the global peak is accessible, meaning that fitness increases monotonically along the path [212], [216], [217]. In contrast to expectation, we observe that the global peaks of landscapes produced under more robust codes are not significantly more accessible for the ParD-ParE2 landscape ($R = 0.0052$, $p = 0.099$), and significantly less accessible for the ParD-ParE3 landscape ($R = -0.151$, $p < 2.2 \cdot 10^{-16}$).

We reasoned that the accessibility of the global peak might be confounded by its size: As the number of codons encoding an amino acid ranges from 1 to 6 in the amino acid permutation codes, the number of distinct mRNAs encoding a given protein variant can range from 1 to 6^L, where L is the number of sites in the protein. Indeed, we observe that the mutational accessibility of the global peak is strongly correlated with its size ($R = 0.801$, $p < 2.2 \cdot 10^{-16}$ for GB1; Supp. Fig. S4.4). Moreover, due to the fact that one of the synonymous codon blocks is split (UCN and AGY, with N denoting any nucleotide and Y denoting U or C; encoding serine in the standard

genetic code), there might be several disconnected regions of the landscape encoding the protein sequence with the highest fitness value. When this is the case, the mutational accessibility of the global peak is significantly higher compared to codes where the global peak comprises a single connected region in genotype space (e.g., in the GB1 landscape, the mutational accessibility is 0.086 vs. 0.055, $p < 2.2 \cdot 10^{-16}$, Welch two sample t-test; Supp. Fig. S4.5). In order to make the landscapes more comparable, we restricted our analysis to only those landscapes in which the size of the global peak, in terms of number of mRNAs encoding the corresponding protein sequence, is the same as in the standard genetic code and, moreover, none of the amino acids contained in the global peak sequence are encoded by the split codon block. In this subset of landscapes, we observe the expected positive correlation between code robustness and accessibility of the global peak for all landscapes except for ParD-ParE3 (Fig. 4.2C and Supp. Tab. S4.2).

While statistically significant, the strength of the correlation between code robustness and global peak accessibility is weak and the magnitude of the effect is not large (mean global peak accessibility 0.055 vs. 0.058 for the 1% least and most robust codes, respectively; Fig. 4.2C). Given the low prevalence of sign epistatic interactions in the landscapes generated under even the least robust genetic codes, we reasoned that the range of landscape ruggedness observed in our data may simply be too small to observe a strong positive correlation between global peak accessibility and code robustness. This is indeed the case, as we confirmed by artificially inflating the ruggedness of the GB1 landscape (Supp. Section S4.3).

In sum, the mutational accessibility of the global peak is strongly influenced by the number of its constituent mRNA sequences and whether they occupy disjoint regions of genotype space, and only weakly influenced by code robustness, due to the limited range of landscape ruggedness produced by the 100,000 amino acid permutation codes.

Sensitivity analyses

Our measure of code robustness assigns amino acids to discrete groups based on seven key physicochemical properties, such as whether the amino acids are acidic or basic (Supp. Fig. S4.3; Methods). However, there are hundreds of physicochemical properties that can be used to characterize amino acids and the relevant amino acid properties can be protein- and site-specific. We thus repeated the analyses described above with code robustness defined in terms of each of the 553 properties from the AAindex database [218], [219]. We show that amino acid properties generally either (1) do not influence landscape ruggedness significantly, or (2) influence it in the direction consistent with our previous results, i.e., increased robustness leads to decreased landscape ruggedness (Supp. Section S4.4). We also see that the significant amino acid properties differ systematically among data sets (Supp. Section S4.4).

To more clearly demonstrate the link between fitness-preserving mutations and landscape ruggedness, we also considered an alternative definition of code robustness, calculated as the expected fitness change upon mutation in a particular data set (Supp. Section S4.5). Under this definition, robustness is not characterized by any particular physicochemical property of amino acids or by a combination of several such properties, but rather by which amino acids are interchangeable in a given data set. We again show that, when using this definition of robustness, robust codes lead on average to smoother fitness landscapes, with markedly stronger correlations than when using the aggregate measure of code robustness (Supp. Section S4.5).

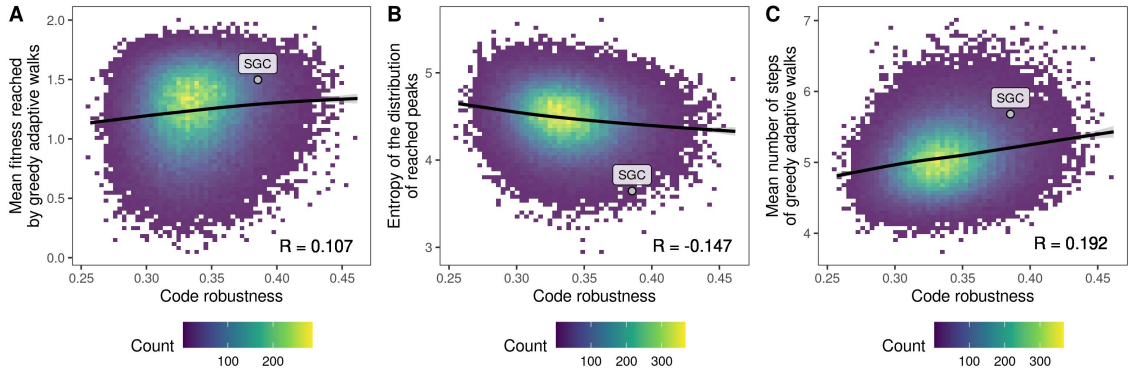


FIGURE 4.3: Relationship between code robustness and results of greedy adaptive walks. The labeled point denotes the results obtained using the standard genetic code. Data pertain to GB1.

Finally, we show that our results are qualitatively insensitive to the way the rewired genetic codes are generated (Supp. Sections S4.6 and S4.7)) and to the dimensionality of the data set (i.e., the number of protein sites; Supp. Section S4.8).

4.3.3 *Evolutionary simulations reveal complex relationship between code robustness and evolvability*

Our analyses suggest that code robustness promotes evolvability by producing smooth adaptive landscapes with few peaks and little sign epistasis. As a consequence, we anticipate evolving populations to obtain higher fitness, on average, when translating proteins using more robust codes than when using less robust codes. To determine if this is the case, we turn to evolutionary simulations, specifically of greedy adaptive walks [181]. These model adaptive evolution of a large population with pervasive clonal interference, such that all possible point mutations to a sequence are simultaneously present in the population, and the fittest of these variants goes to fixation. For each of the 100,000 amino acid permutation landscapes and each of the six data sets, we initialized the walks in each nucleotide sequence encoding a functional product. We terminated a walk when it reached a local or global adaptive peak, and recorded the fitness of that peak sequence (Methods).

Fig. 4.3A shows the average fitness reached by the greedy adaptive walks in relation to code robustness. As expected from our landscape-based analyses, evolving populations reached higher fitness, on average, when translating proteins using more robust genetic codes for the GB1, ParD-ParE2, and ParB-*NBS* landscapes (Supp. Tab. S4.4). However, the results for the ParD-ParE3 landscape were not statistically significant ($R = -0.004$, $p = 0.182$) and we even observed a negative correlation in the ParB-*parS* ($R = -0.0118$, $p = 1.89 \cdot 10^{-4}$) and DHFR data sets ($R = -0.096$, $p < 2.2 \cdot 10^{-16}$). Similar to the analysis of accessible paths, we reasoned that these results might be caused by variation in the size of the global peak, such that larger global peaks are easier to "find" than smaller global peaks, simply because they contain more mRNA sequences. Indeed, we observe a positive correlation between the size of the global peak and mean fitness reached by the greedy adaptive walks in all six data sets (Supp. Tab. S4.5). We thus again restricted our analysis to those genetic codes for which the size of the global peak is the same as under the standard genetic code and occupies a single connected region in genotype space. However, even in

this subset of codes, we observe a positive correlation between code robustness and mean fitness reached by the greedy adaptive walks in only 4 out of 6 data sets (Supp. Tab. S4.4). Whether code robustness promotes or diminishes evolvability thus depends on the particular landscape, which is surprising given that robust genetic codes are associated with smoother fitness landscapes in all 6 data sets. In Supp. Section S4.9, we show that the correlations between code robustness and mean fitness result from a complex interplay between the heights and sizes of the basins of attraction of the peaks, as well as from idiosyncracies specific to particular data sets.

We also observe that the average length of the walks tended to be longer under robust codes (Fig. 4.3C; 4.89 vs. 5.30 steps, on average, for the 1% least and most robust codes, respectively), revealing that the benefit of increased fitness afforded by code robustness comes at the cost of longer evolutionary trajectories to adaptation. This is in line with our observations concerning landscape ruggedness. In landscapes with many local peaks, a greedy walk is more likely to be initialized near one of these peaks, which it will likely ascend in only a small number of mutational steps. In contrast, in landscapes with few local peaks, a greedy walk is more likely to be initialized farther away from one of these peaks, thus increasing the length of the mutational path to adaptation, be it to a local or global peak.

We also highlight that the greedy walks reached higher mean fitness under the standard genetic code than under 85% of the amino acid permutation codes in five out of six landscapes and ranked among the top 5% in three of them. Similarly, the standard code resulted in exceptionally low Shannon entropy of the distribution of reached peaks in five out of six data sets (Supp. Tab. S4.3), meaning that under the standard genetic code, greedy walks preferentially converged on a small number of fitness peaks. We observe qualitatively the same results in simulations of the weak-mutation regime (Supp. Section S4.10) and using codes constructed by restricted amino acid permutation (Supp. Section S4.6) and random codon assignment (Supp. Section S4.7).

4.3.4 *The genetic code governs the genetic architecture of long-term molecular evolution*

In the previous section, we studied a short-term adaptive process, in which high-fitness protein variants evolve from low-fitness variants via mutation and selection. However, once an evolving population reaches high fitness, it behaves like a random walk amongst the mutationally-interconnected set of high-fitness variants. To assess how different code rewirings influence this random walk, we apply a visualization technique that captures the dynamics of a finite population evolving on a fitness landscape at mutation-selection-drift balance [220] where the distances between genotypes reflect the expected amount of time to evolve form one genotype to another (squared distances have units of time, and time is scaled such that each nucleotide mutation occurs at rate 1, see Methods).

In an earlier study, we used this technique to explore the structure of the GB1 landscape at the amino acid level [221] and found that it consists of three main regions of high-fitness protein variants that differ primarily in the placements of small non-polar and bulkier amino acids at positions 41 and 54 (Supp. Fig. S4.6). The first and largest of these regions is characterized by having G at position 41, which is compatible with most amino acids at position 54 and contains the wild-type sequence (VDGV); we will refer to this as Region 1. The second region is characterized by A at position 54, which can be paired at position 41 with C, S, A, and, to a lesser extent, L and

F, and we will refer to this as Region 2. The final region, Region 3, typically has G at position 54, while tolerating T at 54 in some contexts, together with L or F at position 41. Moreover, each of these three regions is connected via functional intermediates with the other two regions (see Supp. Fig. S4.6 and ref. [221]).

Here, we consider how the genetic code, standard or otherwise, reshapes the structure of these regions and restricts their mutational interconnections, focusing on the standard genetic code as well as the two most and two least robust in our set of 100,000 amino acid permutation codes (see Supp. Fig. S4.7 for the corresponding codon tables). Fig. 4.4 shows the resulting visualizations, where for each code we plot the visualization using a sufficient number of dimensions to show the major features of the corresponding fitness landscape. These dimensions are called Diffusion Axes because they reflect the dynamics of diffusion in sequence space, and they are ordered such that the first k Diffusion Axes provide an optimal approximation of the expected times to evolve from one sequence to another (see Methods). In order to better see the structure of the high fitness set, we also show a second visualization for each code where we only plot the high fitness sequences (which in what follows we will take to be the fittest 1%) and color these sequences by their corresponding region in amino acid sequence space.

We find that different rewirings of the genetic code produce fitness landscapes with dramatically different structures from each other or from the structure of the fitness landscape in amino acid sequences space. For example, under the standard genetic code, Region 1 is no longer directly connected to Region 3 because neither 41F nor 41L is accessible from 41G under the standard genetic code (Fig. 4.4A). Indeed, under many codes the set of high fitness sequences becomes split into several distinct components separated by lower fitness sequences, as observed in Robust Code A (Fig. 4.4B), Robust Code B (Supp. Fig. S4.8), and Non-Robust Code B (Fig. 4.4E), so that moving from one component to another requires the fixation of less fit sequences. The waiting time for such deleterious fixations is long, increasing the amount of time required for a population to explore the landscape. We can quantify this in terms of the relaxation time of the rate matrix for the evolutionary random walk, given by the inverse of the absolute value of its largest non-zero eigenvalue. For Robust Code A and Non-robust Code B, the relaxation time is, respectively, 3.17 and 3.22-fold longer than the expected waiting time for individual nucleotide mutations; this is roughly 50% longer than for the standard genetic code or Non-robust Code A with relaxation times of 2.31 and 2.05-fold longer than the expected waiting time for individual nucleotide mutations (see Methods for details).

We also find that high fitness sequences connected by high fitness paths can be connected via very different structures in sequence space. For example, the standard genetic code, as well as the two least robust codes (Fig. 4.4A, D, and E) all show long branch-like structures where the high-fitness paths connecting genotypes require that the mutations be accumulated in a specific order. This can result in very long paths, for example Supp. Fig. S4.9 shows an example of an 11-mutation path connecting Region 1 and Region 3 under the standard genetic code that does not include any substitutions at positions 39 or 40, synonymous changes, or reversions. In contrast, we can also see regions of sequence space where the high fitness set shows a grid-like structure in which mutations at a pair of sites can accumulate independently from each other, as seen in the right hand panel of Fig. 4.4B for Robust code A, or under the standard genetic code, where a

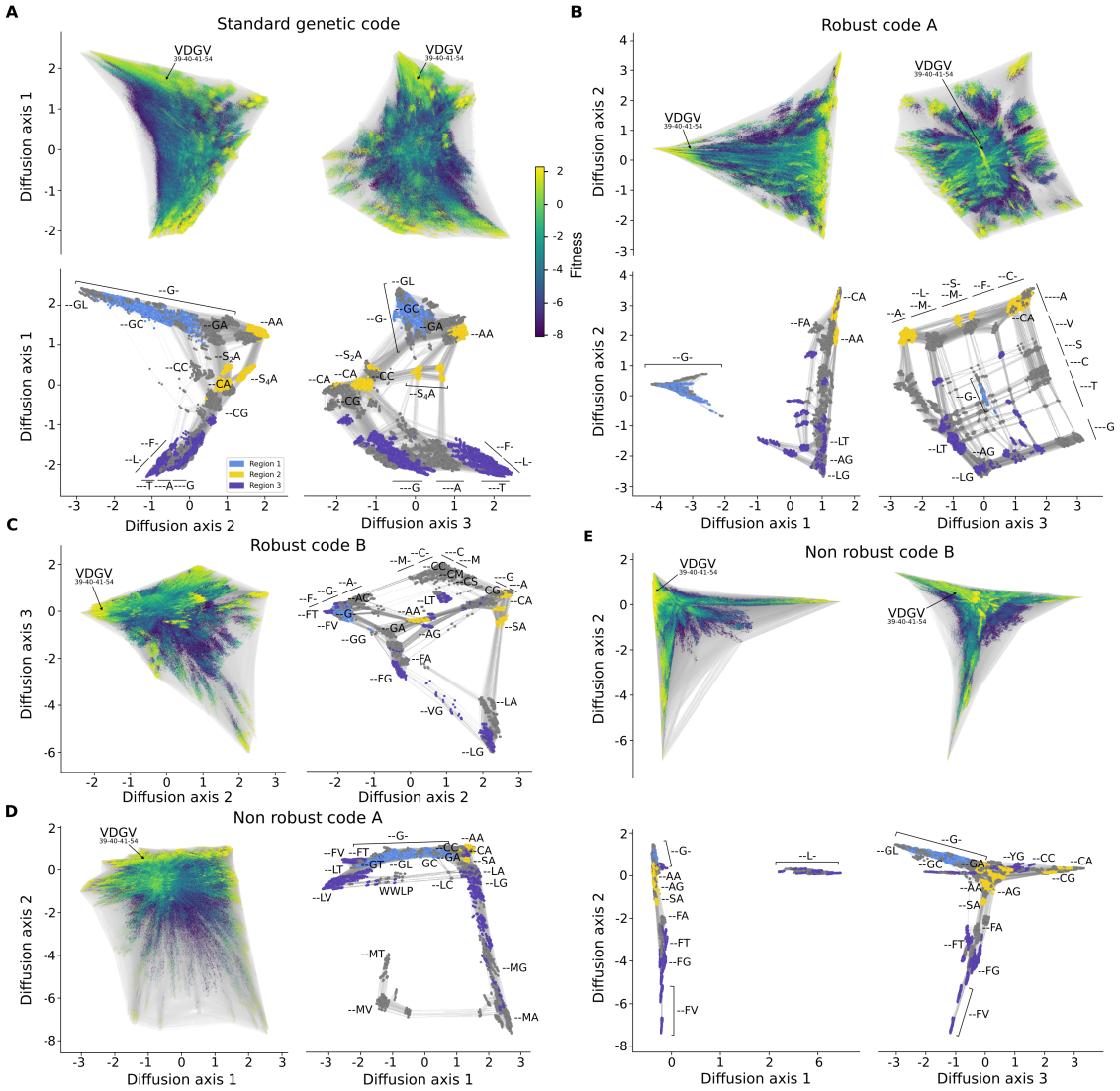


FIGURE 4.4: The genetic code governs genotype network topology and the genetic architecture of long-term molecular evolution. Fitness landscape for GB1 at positions 39, 40, 41 and 54 under the (A) standard genetic code, (B, C) the two most and (D, E) the two least robust codes in the amino acid permutation set. Vertices represent 12-nucleotide sequences and edges connect vertices if their corresponding sequences differ by a single point mutation. Vertex color represents protein fitness (color bar in (A) applies to all panels). Vertices are placed at the coordinates along the diffusion axes, which at a technical level are defined by the subdominant eigenvectors of the rate matrix describing the weak mutation dynamics and have units of square root of time [220], and where time is scaled such that each possible nucleotide mutation occurs at rate 1 (see Methods for details). For each pair of diffusion axes shown, there are two subpanels: one that shows all ≈16 million genotypes, with the location of the sequences encoding the wild-type protein sequence V39 D40 G41 V54 marked, and another that shows only the genotype network of high-fitness variants (top 1% of fitness distribution), which better shows the connectivity between high-fitness regions and which is annotated with the protein sequence features that characterize each cluster or subset of nucleotide sequences. Colors in these panels represent the main regions of functional protein sequences as highlighted in Supp. Fig. S4.6 and show the extent to which the connectivity between these regions of amino acid sequence space is rewired under the different genetic codes.

population can switch between F or L at position 41 more or less independently of whether T, A or G is found at position 54 (Fig. 4.4A, negative values of Diffusion Axis 1).

Besides these large-scale differences in the pattern of connectivity between high-fintess sequences, the density of high-fitness paths can also vary greatly. One particularly interesting case is where a pair of sequences are connected by many long high-fitness paths but are also accessible via a smaller number of short high-fitness paths that can only be accessed on very specific genetic backgrounds; we call these rare shortcuts "wormholes" because they are short paths that connect otherwise distant regions of the high-fitness set. For example, under the standard genetic code, we can see that distant parts of the network of high fitness sequences are in fact accessible from one another via $41S_4$ (where the two disconnected sets of serine codons are broken into S_2 and S_4, named for the number of codons in each set [71]; Fig. 4.4A bottom right). In this case, the average probability for 41G-54L sequences of arriving at a high fitness genotype with L or F at position 41 and T, A or G at position 54 through a high fitness S_4 intermediate is 1.13%. Thus, although these short paths are possible, they occur only a small minority of the time. We see another example of such a wormhole under Non-robust code A, where WWLP sequences bridge the otherwise distant regions characterized by 41L-54A and 41L-54T (Fig. 4.4D). This wormhole is used even more rarely, only 0.002 % of the time.

In summary, these visualizations illustrate the richness and variety of landscape topographies that can be induced by different genetic codes, and the extent to which even exceptionally robust codes can interact with the idiosyncrasies of a particular protein fitness landscape to break crucial links between high-fitness variants.

4.3.5 *Codon compression schemes reveal additional code features influencing evolvability*

Above, we have focused on amino acid permutation codes. However, engineering these codes in a living organism would require an extensive recoding of the genome, including the engineering of many orthogonal aminoacyl tRNA synthetases and tRNAs. For example, in the most robust of the 100,000 codes we analyzed, only two amino acids occupy the same synonymous codon block as in the standard genetic code (Supp. Fig. S4.7B). In contrast, to date, the synthetic biology community has engineered rewired genetic codes that change the meaning of up to only a small handful of codons [40], [41], [222], [223]. There is therefore a large disconnect between the space of theoretically- and practically-realizable rewired genetic codes.

This motivated us to study a subset of rewired genetic codes that require only a small number of codon reassignments, as it may be possible to engineer these codes in a living organism using currently available technology. In particular, we studied the 57-codon *E. coli* genome synthesized by Ostrov et al. [187], in which all occurrences of 7 codons, from 4 synonymous codon blocks, together with the corresponding tRNAs were removed from the genome and are thus theoretically free for reassignment (Supp. Fig. S4.10). Assuming each of the 4 synonymous blocks is reassigned to one amino acid or a stop signal (as might be required by the tRNA wobble rules [224], [225]), there are in total $21^4 = 194,481$ possible rewirings based on this compression scheme, one of them being the standard genetic code. We computationally generated all of these “Ostrov” codes and repeated the landscape-based analyses and evolutionary simulations described above.

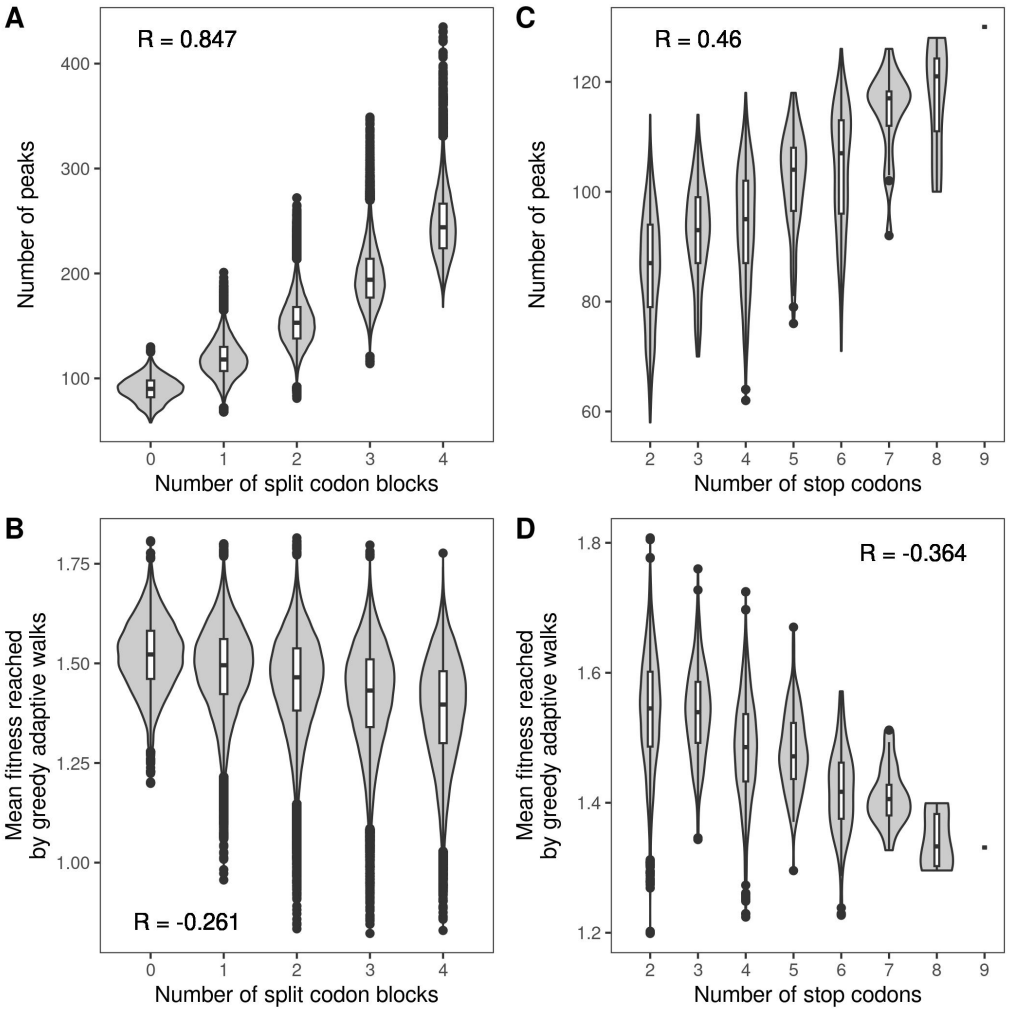


FIGURE 4.5: Additional code features influencing protein evolvability. Violin plots of the number of local peaks and the mean fitness reached by the greedy adaptive walks, shown in relation to (A, B) the number of split codon blocks in the 194,481 Ostrov codes and (C, D) the number of stop codons in the 3,965 Ostrov codes with no split codon blocks. Data pertain to GB1. The violin plots show the distribution and the box-and-whisker plots the median, 25th and 75th percentile. The upper whisker extends from the top of the box to the largest value no further than 1.5-times the inter-quartile range, the lower whisker extends from the bottom of the box to the smallest value no further than 1.5-time the inter-quartile range. Data beyond the end of the whiskers are plotted individually.

Relative to the amino acid permutation codes, the Ostrov codes exhibited even stronger trends with respect to landscape ruggedness (Supp. Tab. S4.6) and the outcomes of the greedy adaptive walks (Supp. Tab. S4.7). Notably, the range of the landscape ruggedness measures, e.g., the number of peaks, is roughly the same as for the amino acid permutation codes, even though the Ostrov codes exhibit a much smaller range of code robustness (from 0.330 to 0.406, as compared to 0.257 to 0.462 for the permutation codes). Because the Ostrov codes differ from the permutation codes in that they do not all have the same synonymous codon block structure or the same number of stop codons as the standard code, we reasoned that these two structural features may provide an explanation for these observations.

The Ostrov codes can have more or fewer split codon blocks than the standard code. In the set of the 194,481 Ostrov codes, the number of split codon blocks ranges from zero to four (Supp. Fig. S4.11). Increasing the number of split codon blocks decreases code robustness ($R = -0.345, p < 2.2 \cdot 10^{-16}$; Supp. Fig. S4.12A), due to the increase in the number of non-synonymous mutations. This causes an increase in landscape ruggedness (Fig. 4.5A and Supp. Tab. S4.8), because maladaptive valleys can form in the mutational spaces between synonymous codons of split codon blocks. Consequently, the average fitness reached by the greedy adaptive walks consistently decreases as the number of split codon blocks increases in all data sets except for DHFR (Fig. 4.5B and Supp. Tab. S4.9). Consistent with the growing number of local peaks, we also observe that the adaptive walks get on average shorter and their endpoints less predictable as the number of split codon blocks increases (Supp. Tab. S4.9).

The Ostrov codes can also have more or fewer stop codons that the standard code, which has three (UAG, UAA, and UGA). Because only the stop codon UAG has been freed for reassignment in the Ostrov codes, the minimum number of stop codons is two, whereas the maximum is nine, corresponding to the assignment of all freed codons to a termination signal (Supp. Fig. S4.13). Supp. Fig. S4.12B shows that increasing the number of stop codons tends to decrease code robustness ($R = -0.160$, $p < 2.2 \cdot 10^{-16}$), due to the increase in the number of nonsense mutations. Moreover, the number of stop codons is negatively correlated with the number of split codon blocks ($R = -0.269$, $p < 2.2 \cdot 10^{-16}$), because if a codon block is assigned to a stop signal, it cannot be part of a split codon block. Thus, when measuring the effect of the number of stop codons on landscape ruggedness or the outcomes of adaptive walks, one has to condition on a given number of split codon blocks. In the following, we report results for codes with no split codon blocks; results for other numbers of split codon blocks can be found in Supp. Tab. S4.10 and S4.11. We observe that increasing the number of stop codons leads to an increase in the number of local peaks (Fig. 4.5C and Supp. Tab. S4.10), as well as decreased accessibility of the global peak (Supp. Tab. S4.10); the effect on epistasis is more complex (Supp. Tab. S4.10 and Supp. Section S4.11). Correspondingly, the average fitness reached by the greedy adaptive walks decreases as the number of stop codons increases (Fig. 4.5D and Supp. Tab. S4.11). This is expected, as in our adaptive landscapes, sequences containing stop codons are assigned a fitness value lower than any of the sequences without stop codons (Methods), reflecting the fact that the inclusion of a stop codon in an open reading frame causes the premature termination of translation and thus protein truncation, which is usually deleterious to protein function. We also observe that the greedy adaptive walks get shorter and less predictable as the number of stop codons increases (Supp. Tab. S4.11).

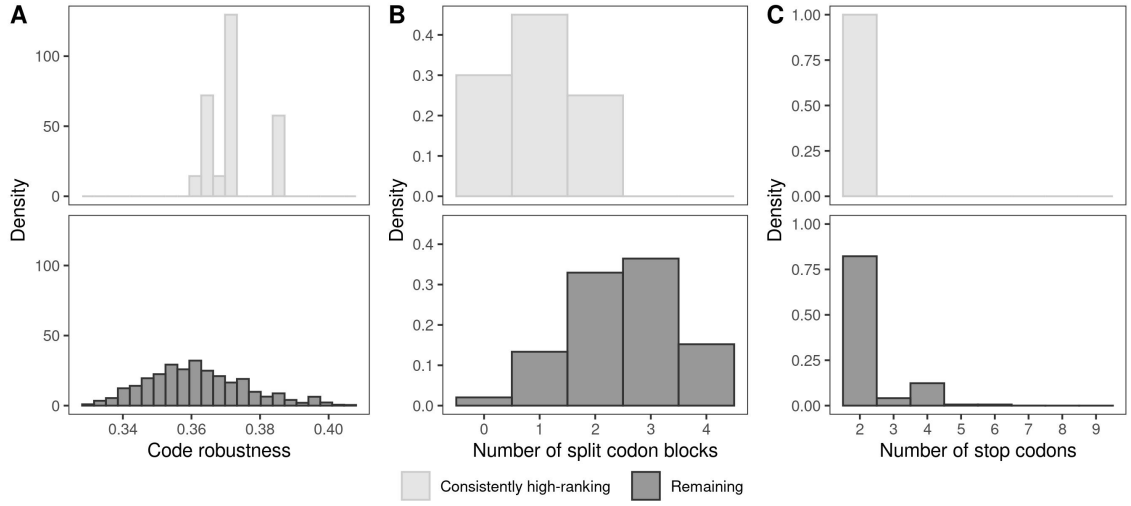


FIGURE 4.6: Design principles for enhancing evolvability. Comparison of the properties of the 20 consistently high-ranking (top 25%) codes (top row) with the remaining 194,461 codes (bottom row), in terms of (A) code robustness, (B) number of split codon blocks, and (C) number of stop codons.

In sum, our analyses of all possible code rewirings under the codon compression scheme proposed by Ostrov et al. [187] reveal additional code features influencing protein evolvability, namely the number of split codon blocks and the number of stop codons.

4.3.6 *Design principles: Genetic codes enhancing and diminishing evolvability*

As discussed in Supp. Sections S4.4 and S4.5, the amino acid properties most relevant for landscape topography are data set-dependent. This suggests that a genetic code that promotes evolvability for one protein might not do so for another. Indeed, in our evolutionary simulations with the Ostrov codes, the mean fitness reached by the greedy walks is not strongly correlated across our six data sets (Supp. Tab. S4.12). Nonetheless, there is a small subset of codes that promote evolvability across all six data sets, and we reasoned that these may exhibit commonalities that could inform design principles for engineering genetic codes to promote evolvability across a diversity of proteins.

We therefore ranked each Ostrov code in descending order according to mean fitness reached in the evolutionary simulations, separately for each of the six data sets. There are 20 codes that consistently rank in the top 25% of all six lists. We note that the standard genetic code is not a member of this set of consistently high-ranking codes, as it does not rank in the top 25% of codes for the GB1, ParD-ParE3, and DHFR data sets (Supp. Tab. S4.13). This shows that even relatively small changes to the standard genetic code can enhance protein evolvability.

We then compared these consistently high-ranking codes to the remaining 194,461 codes, in terms of robustness, number of split codon blocks, and number of stop codons. The consistently high-ranking codes have significantly higher robustness ($p = 1.06 \cdot 10^{-5}$, Welch two-sample t-test; Fig. 4.6A), fewer split codon blocks ($p = 2.35 \cdot 10^{-8}$, Welch two-sample t-test; Fig. 4.6B), and fewer stop codons ($p < 2.2 \cdot 10^{-16}$, Welch two-sample t-test; Fig. 4.6C). Similarly, for the 280 genetic codes that consistently rank among the bottom 25% of codes, we observe significantly lower

robustness ($p < 2.2 \cdot 10^{-16}$, Welch two-sample t-test) and more split codon blocks ($p < 2.2 \cdot 10^{-16}$, Welch two-sample t-test) compared to the remaining codes; however, the consistently low-ranking codes do not tend to have more stop codons ($p = 1$, Welch two-sample t-test) (Supp. Fig. S4.14). This suggests that there are some basic design principles to engineering genetic codes that promote evolvability across a diversity of proteins. Specifically, minimize the number of split codon blocks, minimize the number of stop codons, and assign amino acids to codon blocks such that point mutations cause only small changes to amino acid properties, using an aggregate measure of a diversity of amino acid properties [93]. The opposite is true for codes that diminish evolvability, perhaps with the exception of the number of stop codons. To illustrate these design principles, in Supp. Section S4.12 we give examples of Ostrov codes that, in our simulations, consistently decrease or increase evolvability, and compare them with codes specifically designed to promote [93] or diminish [195] evolvability.

4.4 DISCUSSION

The standard genetic code defines the rules of protein synthesis for nearly every life form on Earth [1]. It imparts an extreme, "one in a million" level of error tolerance [10] that buffers the deleterious effects of infidelity in replication, transcription, and translation [9], [10], and provides a striking example of biological robustness at the heart of an essential cellular information processing system [226]. However, prior theoretical work, limited by a lack of suitable data, has disagreed on whether such robustness hinders [93] or facilitates [185] protein evolvability. Here, by computationally translating millions of mRNA sequences under hundreds of thousands of rewired genetic codes using experimental data for six proteins, we reveal that code robustness facilitates protein evolvability by rendering smooth adaptive landscapes upon which evolving populations readily find mutational paths to adaptation. At the same time, our results suggest that the standard genetic code is likely not "one in a million" with respect to evolvability. Moreover, whereas the correlations we observe between code robustness and landscape ruggedness are consistent in direction across data sets, they are relatively weak, so that genetic codes with a similar degree of robustness may differ substantially in the degree of evolvability they confer.

Landscape ruggedness has long been viewed as an impediment to adaptation [176] and, as such, has been used as a proxy for evolvability [178]. The intuition is that ruggedness frustrates evolvability by blocking "uphill" mutational paths to the global adaptive peak, thus limiting the ability of mutation to bring forth adaptive phenotypic variation. Our results confirm this intuition in the context of rewired genetic codes, in that adaptive walks tend to achieve higher fitness on smooth landscapes caused by robust codes than on rugged landscapes caused by less robust codes. However, this is not solely attributable to an increase in the mutational accessibility of the global peak. Rather, as ruggedness decreases, a positive correlation emerges between the height of a peak and the size of its basin of attraction. This causes adaptive walks to preferentially converge on a small number of high-fitness peaks in less rugged landscapes, and more uniformly to all peaks in more rugged landscapes. Moreover, as illustrated by the DHFR data set (Supp. Section S4.9.1), smooth landscapes caused by robust codes may comprise fitness peaks that are lower than the fitness peaks of rugged landscapes caused by non-robust codes, such that increased landscape ruggedness is actually associated with increased protein evolvability. As such, simplistic measures

of landscape ruggedness based solely on the number of peaks may be an insufficient proxy for evolvability [178] or for predicting evolutionary dynamics [227].

An additional factor influencing the accessibility of the global adaptive peak is its size. By randomly permuting amino acids amongst synonymous codon blocks, we created landscapes that vary significantly in the number of mRNA sequences that translate to the highest-fitness protein sequence. In comparing the outcomes of evolutionary simulations on these landscapes, we observe that protein sequences encoded by a large number of mRNA sequences are easier to evolve than equally fit sequences encoded by fewer mRNA sequences [228]–[230]. This observation implies that amino acids encoded by a large number of codons, such as serine or leucine, should be relatively more abundant in protein sequences than amino acids encoded by few codons, such as methionine or tryptophan. Indeed, across the tree of life, there is a positive correlation between the abundance of an amino acid and its number of constituent codons [30], [231], and as early as 1973, Jack L. King attributed this correlation to differences in amino acid "findabilities" caused by the structure of the standard genetic code [232]. Our results generalize this observation to non-standard genetic codes, and suggest that if life had converged on a different standard code, the amino acid composition of proteins would likely be very different from the one we know. Such variation in the proteomic abundance of amino acids may already be apparent in the proteomes of organelles and organisms that use non-standard genetic codes in nature [1], [43], [59], and may emerge in directed laboratory evolution experiments that use synthetic organisms with non-standard genetic codes [40], [188]. If so, these systems may provide empirical support for entropic arguments regarding adaptation [229], [230].

There are several caveats to the results presented here. First, because there are so few combinatorially-complete data sets measuring a quantitative phenotype for all 20^L protein variants, our conclusions are based on only six empirical adaptive landscapes. Moreover, the six landscapes differ in the strength of the relationship between code robustness and landscape ruggedness (see, e.g., Supp. Tab. S4.2) and, in some cases, even in the sign of the trend. We believe this results from the low dimensionality of our landscapes, in that the landscapes are largely influenced by idiosyncracies specific to a single protein site [233], [234]. For example, we observe that the fitness reached by greedy walks in the DHFR landscape is mostly determined by whether the genetic code allows a mutation from cysteine to aspartate or glutamate (Supp. Section S4.9.1). Indeed, our ablation analysis of the GB1 landscape suggests that the average strength of the correlation between robustness and evolvability increases as the number of protein sites increases from 2 to 4 (Supp. Section S4.8). Moreover, the data sets we use differ in several important aspects – the level of ruggedness of the corresponding landscape under the standard code, the location of the screened residues in the protein, as well as the assayed phenotype – all of which support the generality of our findings. Second, for three of our data sets, the measured phenotype is the relative binding affinity of the protein to its ligand, and it is not immediately apparent how this phenotype relates to organismal fitness. Even so, a large body of literature attests to the power of such quantitative phenotypes in teaching us about protein evolvability [209]. Third, in the main text, we defined code robustness using a discrete categorization of amino acids based on their physicochemical properties (Supp. Fig. S4.3). A mutation between two amino acids belonging to the same physicochemical group was considered "synonymous". While it is well known that some amino acid pairs are more exchangeable than others [235]–[237], it is a simplification to

assume that these amino acids are entirely equivalent. To overcome this limitation, we additionally considered definitions of code robustness based on a large number of diverse physicochemical properties, one at a time (Supp. Section S4.4), and on a custom amino acid similarity score specific to each of the six data sets (Supp. Section S4.5), with qualitatively the same results. Fourth, the error tolerance of a genetic code, standard or non-standard, is influenced by mutation bias and codon usage [10], [238] as they make some mutations more likely than others. While mutation bias and codon usage may influence peak accessibility in adaptive landscapes [239], they do not affect landscape topography, which is why we have not considered these effects here. We hope that in the future it will become possible to overcome these caveats and confirm our results, both theoretically, as more and larger combinatorially-complete data sets become available, and experimentally, by comparing the dynamics and outcomes of laboratory evolution experiments with proteins and organisms that use different genetic codes.

Such experiments are becoming more broadly accessible, as a diversity of recoded organisms and plasmid-borne orthogonal translation systems are now commercially available. Moreover, these experiments are becoming increasingly scalable. For example, Zürcher et al. [40] have recently engineered bacterial strains with as many as 16 different genetic codes. Understanding the relationship between code structure and evolvability is therefore highly topical, as the future in which synthetic organisms with non-standard genetic codes are utilized in science and in industry [240], [241], for example to accelerate directed evolution experiments [40], [93], [177], to achieve bio-containment [40], [41], [195], [242]–[244], or to produce drugs [245]–[247] is now tangibly close. We have identified general design principles, as well as a few concrete candidate codes, that are expected to increase evolvability beyond that of the standard genetic code. We have identified even more genetic codes that decrease evolvability, which may be useful for the bio-containment of synthetic organisms. All of these codes are compatible with the 57-codon *E. coli* genome reported by Ostrov et al. [187], and could thus be engineered in the lab using existing technology. Our analyses with this codon compression scheme explored all 194,481 genetic codes that reassign one or more of the freed codon blocks, assuming that the whole synonymous codon block needs to be assigned to one amino acid. If this assumption is lifted [40], [248], it would, in the context of the 57-codon *E. coli* genome, lead to a staggering $21^7 \approx 1.8 \cdot 10^9$ possible code rewirings. Together with other codon compression schemes [189], [249]–[251], the space of possible code rewirings available today is practically infinite, and will continue to grow as larger-scale rewirings become feasible.

Building upon our results, there are several directions for future research. First, the advances in biotechnology discussed above now enable experimental tests of the relationship between genetic code robustness and evolvability, as recently proposed by Zürcher et al. [40]. Do robust genetic codes indeed lead to larger improvements of protein function in directed evolution experiments? And are organisms with robust genetic codes better able to adapt to changing environmental conditions? Second, in this paper we have worked only with rewired genetic codes, i.e., codes that change the mapping between codons and amino acids, but we have not considered expanded genetic codes, i.e., codes that include a 21st, non-standard amino acid. While a small number of evolutionary experiments using organisms with expanded genetic codes have been reported [252]–[255], how the addition of a 21st non-standard amino acid influences protein and organismal evolvability is not yet fully understood. This question could be addressed

experimentally by generating combinatorially-complete data for all 21^L sequence variants, using a diversity of 21st non-standard amino acids, and also theoretically, for example by subsampling combinatorially-complete data to contain fewer than 20 amino acids [256].

4.5 METHODS

Data processing

We estimated the fitness of each measured amino acid variant following Rubin et al. [257].

For each replicate experiment i, the fitness of variant v is equal to

$$f_{v,i} = \log\left(\frac{c_{v,i,\text{sel}} + \frac{1}{2}}{c_{wt,i,\text{sel}} + \frac{1}{2}}\right) - \log\left(\frac{c_{v,i,\text{inp}} + \frac{1}{2}}{c_{wt,i,\text{inp}} + \frac{1}{2}}\right),$$

where $c_{v,i,\text{sel}}$ is the count of variant v in the i-th replicate sample after selection, $c_{wt,i,\text{sel}}$ is the count of the wild type in the i-th replicate sample after selection, $c_{v,i,\text{inp}}$ is the count of variant v in the i-th replicate input sample, and $c_{wt,i,\text{inp}}$ is the count of the wild type in the i-th replicate input sample. The variance of the estimate is equal to

$$\sigma^2_{v,i} = \frac{1}{c_{v,i,\text{inp}} + \frac{1}{2}} + \frac{1}{c_{v,i,\text{sel}} + \frac{1}{2}} + \frac{1}{c_{wt,i,\text{inp}} + \frac{1}{2}} + \frac{1}{c_{wt,i,\text{sel}} + \frac{1}{2}}.$$

The final fitness of variant v is then the weighted average of the r replicates, with weights given by the inverse of the corresponding variance:

$$f_v = \frac{\sum_{i=1}^{r} \frac{1}{\sigma^2_{v,i}} f_{v,i}}{\sum_{i=1}^{r} \frac{1}{\sigma^2_{v,i}}}$$

and the variance is computed as

$$\sigma^2_v = \frac{1}{\sum_{i=1}^{r} \frac{1}{\sigma^2_{v,i}}}.$$

The number of replicates r equals 1 for GB1, ParB-*parS*, and ParB-*NBS*; 2 for ParD-ParE2 and ParD-ParE3; and 6 for DHFR.

In the ParB-*parS* and ParB-*NBS* data sets, read counts of 79,187 variants were not reported in the original publication. It is those variants that had 0 reads in both post-selection samples. When computing the fitness and variance for these variants, we assumed the input read counts are equal to the median input read count over all protein variants with 0 reads after selection for binding a given DNA-motif.

The DHFR landscape was measured on the level of codons, i.e., it contains a fitness measurement for each synonymous encoding of a given protein variant. Because any connection between the original codons and the amino acids is lost in our randomized genetic codes, we assume, as for the other landscapes, that all synonymous variants have the same fitness. This protein-level fitness is computed by first pooling the reads based on the encoded amino acids and then proceeding as described above. In the original study, the bottom 93% of the nucleotide variants, according to

fitness, were considered non-functional [201]; we assume the same for the 7,173 corresponding protein variants (89.7% of the landscape). We assigned these variants the same fitness as variants containing stop codons (see below).

Based on these raw fitness estimates and variances for observed variants, we imputed the fitness values for the missing variants and inferred the full adaptive landscape as the maximum *a posteriori* estimate under empirical variance component regression, an empirical Bayes modeling framework that naturally incorporates all orders of genetic interaction [210].

Constructing adaptive landscapes

To construct an adaptive landscape, we consider the set of all possible mRNA sequences of length 12 (for GB1, ParB-*parS*, and ParB-*NBS*, so that they encode 4 amino acids; there is $4^{12} = 16,777,216$ such sequences) or 9 (for ParD-ParE2, ParD-ParE3, and DHFR, encoding 3 amino acids; $4^9 = 262,144$ sequences), respectively. We represent each sequence as a vertex in a mutational network. Two vertices are connected with an edge if the Hamming distance of the corresponding mRNA sequences is 1, i.e., the two sequences differ by a single point mutation. This underlying network is the same for all genetic codes.

For each genetic code, we then assign an "elevation " to each vertex, equal to the fitness of the sequence, translated using a given genetic code. Sequences containing stop codons are assigned an arbitrary elevation lower than the fitness of any sequence not containing stop codons (we used a value of -100, but the precise value is not relevant for the analyses presented here).

Generating rewired genetic codes

The amino acid permutation codes were generated by randomly permuting the twenty amino acids amongst the synonymous codon blocks. The number and position of stop codons were fixed as in the standard genetic code, as well as the presence of exactly one split codon block (i.e., the UCN and AGY codons, encoding serine in the standard genetic code, always encode the same amino acid). There is $20! \approx 2.4 \cdot 10^{18}$ such codes, of which we randomly sampled 100,000.

For the Ostrov codes, we assumed that each of the freed synonymous codon blocks (UUA+UUG; UAG; AGU+AGC; AGA+AGG) is assigned one of the twenty amino acids or a stop signal. All other codons retain their meaning as in the standard genetic code. In total there is $21^4 = 194,481$ such codes, one of them being the standard genetic code. We generated all of these genetic codes.

Code robustness

We define code robustness as the proportion of single-nucleotide substitutions that do not change the physicochemical properties of amino acids. We divided amino acids into 7 physicochemical groups following Pines et al. [93] (Supp. Fig. S4.3): acidic (D, E); aliphatic (A, I, L, V); aromatic (F, W, Y); basic (H, K, R); glycine (G); polar (C, M, N, Q, S, T); and proline (P). We considered mutations from an amino acid to a stop codon or vice versa as a change in physicochemical properties, whereas we did not consider mutations among stop codons as a change in physicochemical properties.

Number of adaptive peaks

Intuitively, a local peak is a sequence whose fitness is higher than the fitness of any of its neighbors. In our case, due to the degeneracy of the genetic code, several vertices often have the same elevation and, moreover, those vertices will usually be connected; peaks are thus usually plateaus rather than a single vertex. Formally, we define a local peak as a set of vertices that (1) are connected in the genotype space, (2) all have the same elevation, and (3) whose neighbors are either part of the set or have a lower elevation.

Epistasis analysis

A square is a quadruplet of sequences that contains a "wild type" sequence, two of its one-mutant neighbours, and the corresponding double mutant. In the following, we denote by f_{00} the fitness of the wild type, by f_{01} and f_{10} the fitness of the two single mutants, and by f_{11} the fitness of the double mutant. The "mutational effect" of a given mutation is denoted by Δf, e.g., $\Delta f_{00\to10} = f_{10} - f_{00}$ is the change in fitness caused by mutating the wild type sequence to one of the single mutants.

We say that there is no epistasis if

$$f_{00} + f_{11} - f_{01} - f_{10} = 0.$$

The square is classified as having magnitude epistasis if

$$\Delta f_{00\to10} \cdot \Delta f_{01\to11} > 0 \text{ and } \Delta f_{00\to01} \cdot \Delta f_{10\to11} > 0,$$

i.e., the effects of both mutations have the same sign (increase fitness or decrease fitness) regardless of the genetic background. Similarly, the square is classified to have reciprocal sign epistasis if

$$\Delta f_{00\to10} \cdot \Delta f_{01\to11} < 0 \text{ and } \Delta f_{00\to01} \cdot \Delta f_{10\to11} < 0,$$

i.e., the effects of both mutations have opposite signs in different genetic backgrounds. The remaining cases, i.e.,

$$\Delta f_{00\to10} \cdot \Delta f_{01\to11} > 0 \text{ and } \Delta f_{00\to01} \cdot \Delta f_{10\to11} < 0$$

and

$$\Delta f_{00\to10} \cdot \Delta f_{01\to11} < 0 \text{ and } \Delta f_{00\to01} \cdot \Delta f_{10\to11} > 0$$

are classified as simple sign epistasis: the sign of one of the mutations is the same in the different backgrounds, whereas the sign of the second mutation changes in the different backgrounds.

Due to the size of the genotype networks, listing all squares is computationally prohibitive. Thus, we randomly sampled 1,000,000 squares by first sampling a random sequence and then sampling two random mutations at two different positions in the sequence.

Mutational accessibility of the global peak

We define the mutational accessibility of the global peak as the probability that, picking a random functional sequence and a random direct path from the sequence to the global peak, the chosen path is accessible, i.e., the fitness increases monotonically along the path. In our case, the global peak is composed of several mRNA sequences; we define direct paths as those paths that reach any of the global peak sequences in the smallest possible number of steps. For example, considering the standard genetic code and the ParD-ParE3 data set, the global peak consists of 4 sequences: GAU UGG GAA, GAU UGG GAG, GAC UGG GAA, and GAC UGG GAG (all translate to DWE). Starting from sequence GAU UGG AUG (DWM), there are two direct paths to the global peak: GAU UGG AUG - GAU UGG **G**UG - GAU UGG G**A**G and GAU UGG AUG - GAU UGG A**A**G - GAU UGG **G**AG. Notice that in both cases the end point of the direct paths is only one of the 4 global peak sequences, since reaching the other 3 sequences in the global peak would require more than 2 mutations (and hence those paths are not considered direct).

Greedy adaptive walks

We simulated greedy adaptive walks on the landscape in which the most fit of the 1-mutant neighbors is fixed in every step, until a global or local peak is reached. However, the degeneracy of the genetic code means that the fitness values in the landscape are not unique, as all mRNA sequences encoding the same protein share the same fitness. The "most fit" neighbor thus does not have to be uniquely defined, e.g. because there are several possible mutations that lead to the same fitness increase, or because a neutral plateau must be crossed before new adaptive variants may be generated. If this happens, we retain all sequences with the highest fitness; we then explore all of their 1-mutant neighbors and choose the fittest one(s) of those, etc.

We initiated the walks in all possible functional sequences.

Entropy of the distribution of reached peaks

For the greedy adaptive walks, we compute the entropy of the walks' targets as

$$-\sum_{v\in\mathcal{V}} P(v)\log P(v)$$

where $\mathcal{V}$ is the set of all endpoints of the greedy walks (i.e., the set of all adaptive peaks) and $P(v)$ denotes the proportion of greedy walks that terminate on peak v.

Visualization of the GB1 landscape under rewired genetic codes

We used the visualization method as previously described [220]. Briefly, we construct a model of molecular evolution where a population evolves via single nucleotide substitutions and the rate at which each possible substitution becomes fixed in the population is related to its relative

selective advantage or disadvantage. Specifically, the rate of evolution from sequence i to any mutationally adjacent sequence j is given by

$$Q_{ij} = \frac{S_{ij}}{1 - e^{-S_{ij}}}$$

where S_{ij} is the scaled selection coefficient (population size times the selection coefficient of j relative to i) and the total leaving rate from each sequence i is given by

$$Q_{ii} = -\sum_j Q_{ij}.$$

In this context, we assume that the selection coefficient between sequences i and j is proportional to the difference in log-enrichment scores or fitness $f_j - f_i$ and therefore $S_{ij} = c(f_j - f_i)$, where c controls the strength of selection. For all analyses presented here, we used the simple choice of c=1, which for the standard genetic code gives a mean fitness at stationarity equal to 0.055, similar to the wild-type sequence V39 D40 G41 V54 (which in this experiment by definition has fitness 0). Given the rate matrix Q, we then construct the visualization by using the subdominant right eigenvectors r_k associated with the smallest magnitude non-zero eigenvalues λ_k of this rate matrix as coordinates for the low dimensional representation of the landscape, where each such coordinate defines one of the "diffusion axes" used in the visualization. Because the smallest magnitude non-zero eigenvalues and their associated eigenvectors control the most slowly decaying deviations from the stationary distribution, the resulting visualization reflects the long-term barriers to diffusion in sequence space, and clusters in the representation correspond to sets of initial states from which the evolutionary model approaches its stationary distribution in the same way. Thus, multi-peaked fitness landscapes appear as broadly separated clusters with one peak in each cluster. Moreover, by scaling the axes appropriately, as is done here,

$$u_k = \frac{r_k}{\sqrt{-\lambda_k}}$$

these axes u_k have units of square-root of time, where time is measured in the expected number of neutral substitutions for a completely neutral sequence. In particular, using these coordinates u_k, the squared Euclidean distance between arbitrary sequences i and j equals the sum of the expected time H_{ij} to evolve from i to j and the expected time H_{ji} to evolve from j to i, i.e.:

$$\sum_k (u_{k,i} - u_{k,j})^2 = H_{ij} + H_{ji},$$

Using the first several u_k (i.e. u_1 and u_2 for a 2-dimensional representation or u_1, u_2, and u_3 for a three dimensional representation) optimally preserves the above relation in a principal components sense (see ref. [220] for details). Moreover, the associated eigenvalues λ_k reflect the negative rate at which populations diffuse across these barriers, so that $-\frac{1}{\lambda_k}$ gives the timescale it would take to overcome the barriers captured by u_k. Thus, the longest of these timescales, $-\frac{1}{\lambda_1}$ can be used to characterize how fast a population explores a given fitness landscape. Indeed, $-\frac{1}{\lambda_1}$ is also known as the relaxation time of the Markov chain, and is closely linked to many other metrics for describing how quickly a Markov chain approaches its stationary distribution [258].

Additionally, we calculated the probability of using the wormholes under this evolutionary model using standard Markov chain theory [259]. Specifically, we can compute the probability of arriving at a set of genotypes B from another, called A, through specific mutations by defining an absorbing Markov chain in which each mutation entering B is replaced by a corresponding absorbing state, and where the transition rate into this absorbing state is the same rate as the corresponding transition rate in the original chain. We then split these absorbing states into two subclasses, B_1, corresponding to our transitions of interest, and B_2 corresponding to the other transitions into B. We calculate the probability of being absorbed into B_1 for each genotype in A and take the average over all genotypes in A.

Data and code availability

Code used in this study is available at https://github.com/parizkh/rewired_codes_landscapes and all data are publicly available on Zenodo (https://zenodo.org/records/10677993).

ACKNOWLEDGMENTS

We thank Andreas Wagner and Macarena Toll-Riera for discussions, Václav Rozhoň for help with algorithm design and implementation, and Thuy-Lan Lite for kindly providing the counts data needed for the construction of the ParD-ParE2 and ParD-ParE3 landscapes.

4.6 SUPPLEMENTARY MATERIAL

S4.1 *Supplementary figures*

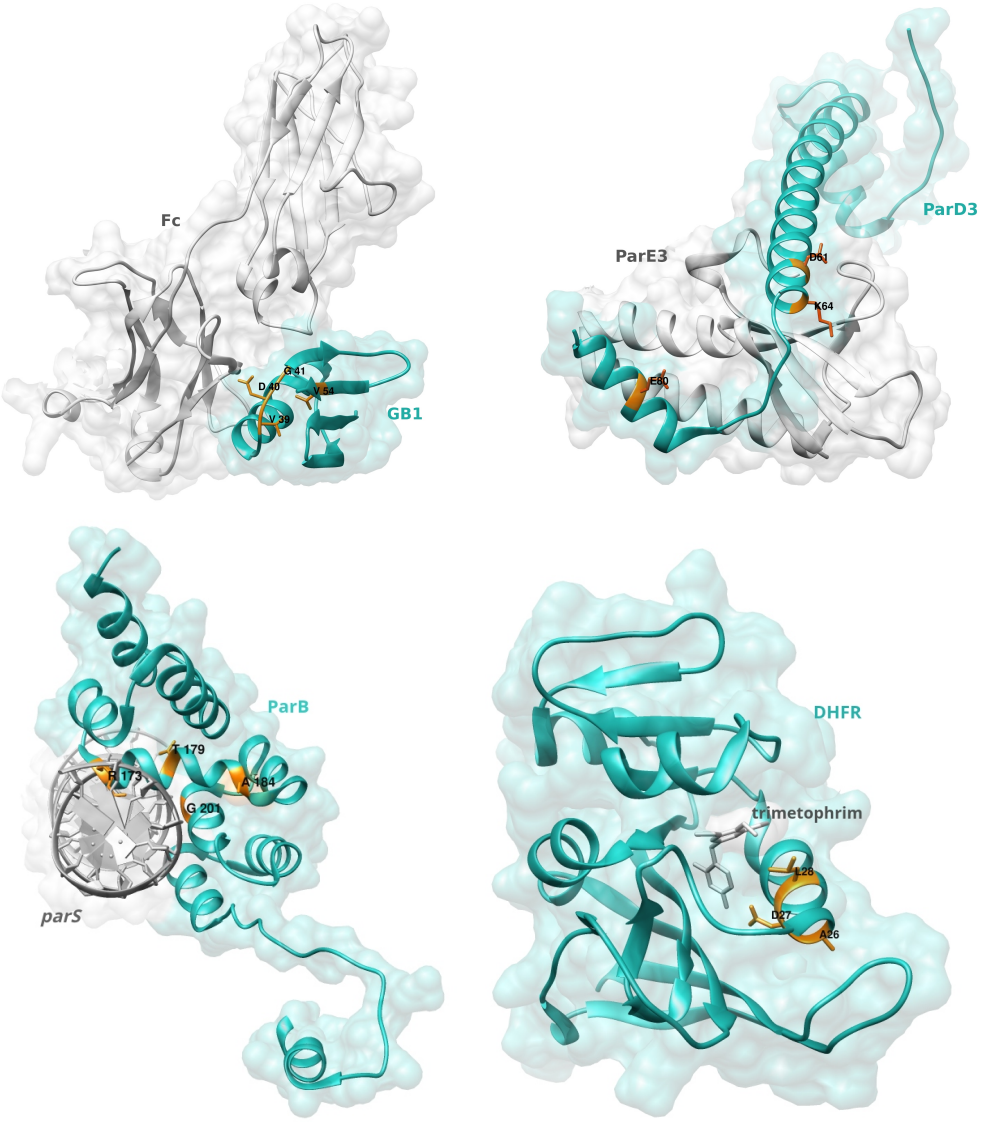


FIGURE S4.1: Structures of the (top left) GB1, (top right) ParD3, (bottom left) ParB, and (bottom right) DHFR proteins, in complex with their corresponding ligands (Fc domain of IgG for GB1; ParE3 for ParD3; *parS* DNA motif for ParB; antibiotic trimetophrim for DHFR). The residues used to build the adaptive landscapes are highlighted in orange. PDB IDs: 1FCC for GB1, 5CEG for ParD3, 6S6H for ParB, 6XG5 for DHFR.

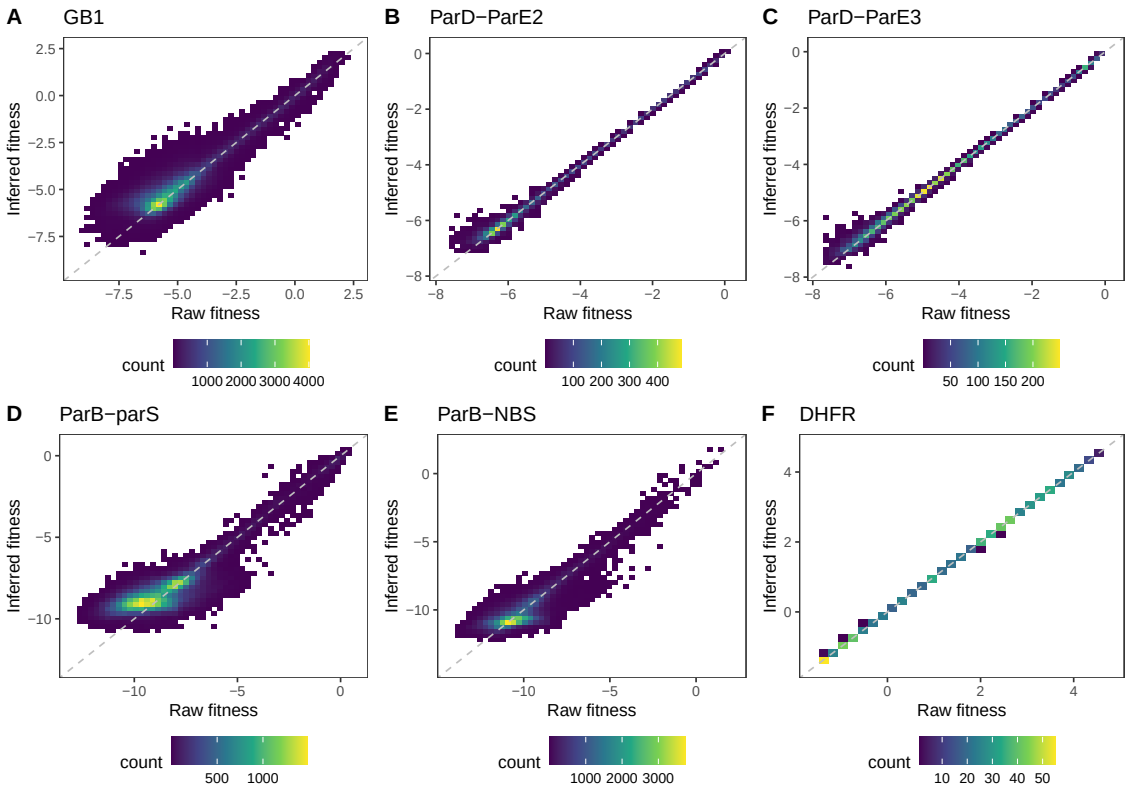


FIGURE S4.2: Density plots of the raw fitness values and the fitness values inferred using empirical variance component regression (Methods) [210] for the (A) GB1, (B) ParD-ParE2, (C) ParD-ParE3, (D) ParB-*parS*, (E) ParB-*NBS*, and (F) DHFR data sets. For DHFR, only the 827 functional variants are shown.

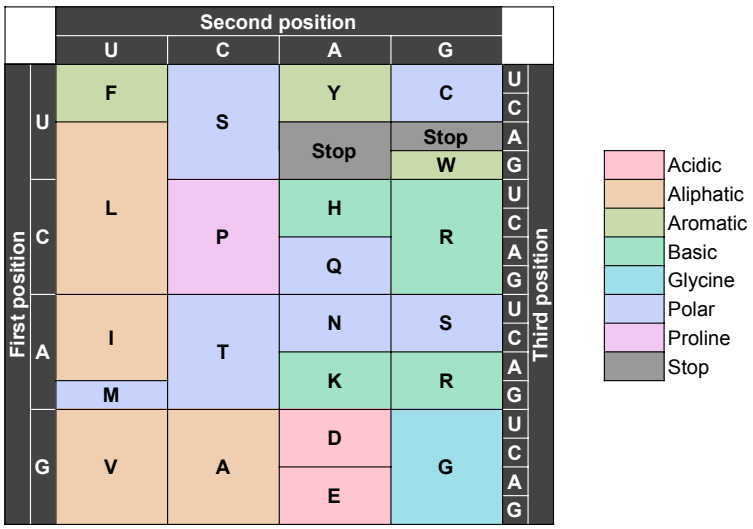


FIGURE S4.3: Codon table for the standard genetic code, with codons colored based on the physicochemical properties of the encoded amino acid. Classification into physicochemical properties taken from ref. [93].

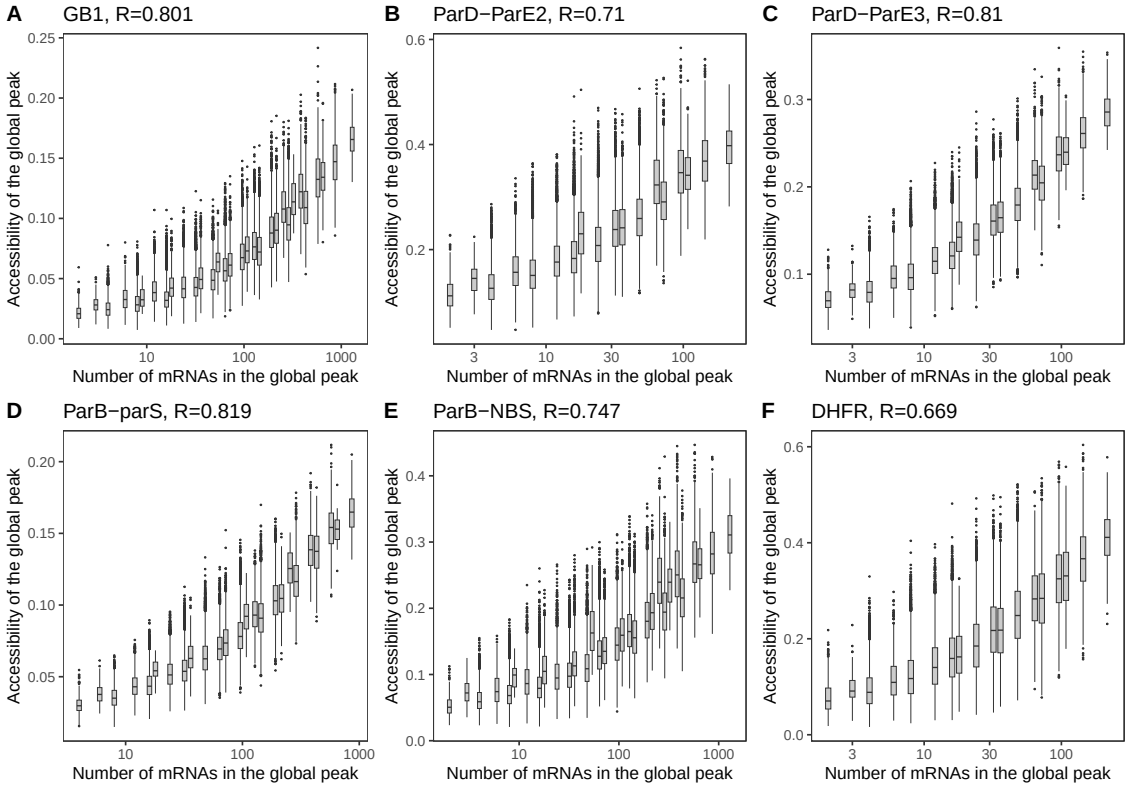


FIGURE S4.4: Accessibility of the global peak in relation to its size for the (A) GB1, (B) ParD-ParE2, (C) ParD-ParE3, (D) ParB-*parS*, (E) ParB-*NBS*, and (F) DHFR landscapes. Mutational accessibility is measured as the proportion of randomly chosen direct paths to the global peak that are accessible, meaning that fitness increases monotonically along the path. Peak size is measured as the number of mRNA sequences encoding the protein with the maximum fitness value. Data pertain to the 100,000 amino acid permutation codes. Values of the Pearson's correlation are shown on the top of each plot. The box-and-whisker plots show the median, 25th and 75th percentile. The upper whisker extends from the top of the box to the largest value no further than 1.5-times the inter-quartile range, the lower whisker extends from the bottom of the box to the smallest value no further than 1.5-time the inter-quartile range. Data beyond the end of the whiskers are plotted individually.

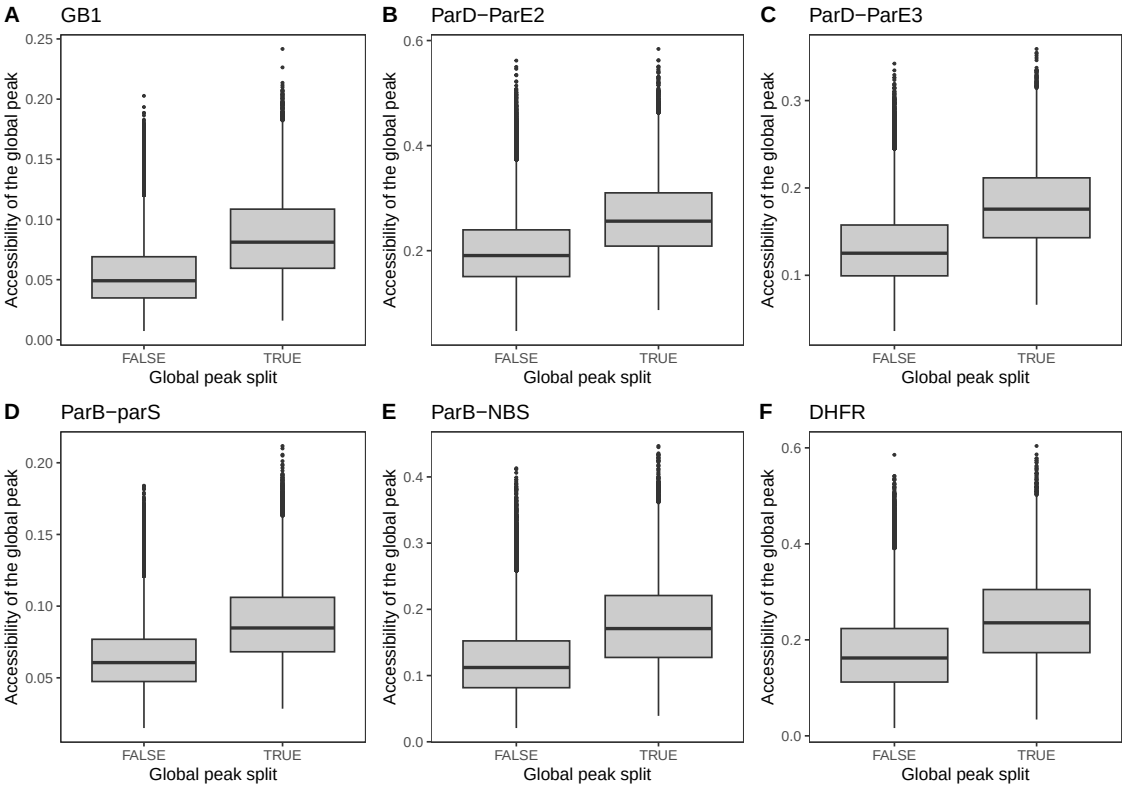


FIGURE S4.5: Accessibility of the global peak in relation to whether it forms a single connected region in genotype space for the (A) GB1, (B) ParD-ParE2, (C) ParD-ParE3, (D) ParB-*parS*, (E) ParB-*NBS*, and (F) DHFR landscapes. The global peak occupies disconnected regions of genotype space when an amino acid in the corresponding protein sequence is encoded by the split codon block. Data pertain to the 100,000 amino acid permutation codes. Meaning of box-and-whisker plots defined in Fig. S4.4.

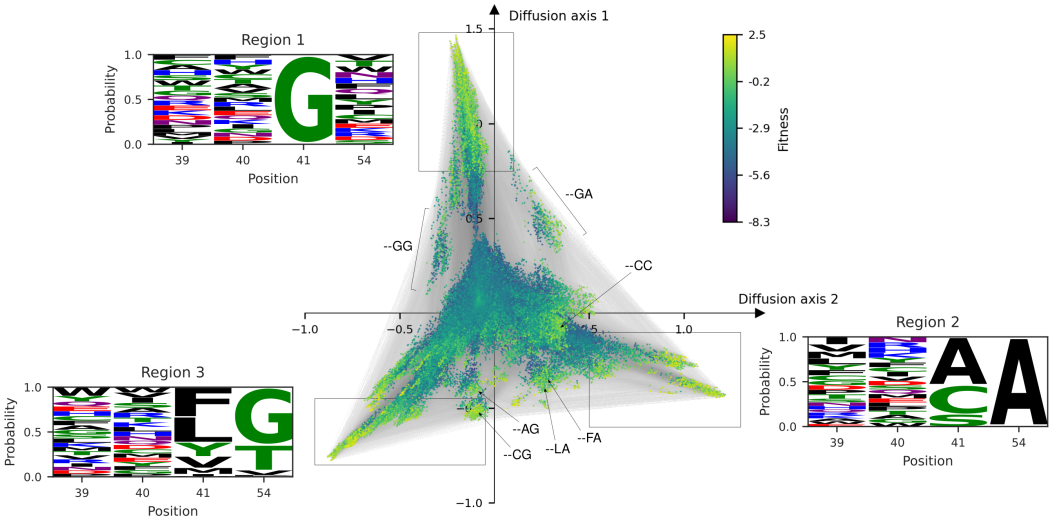


FIGURE S4.6: Visualization of the GB1 fitness landscape at the protein sequence level. Vertices represent 4-amino acid sequences and edges connect vertices if their corresponding sequences differ by a single amino acid substitution. Vertex color represents protein fitness. Vertices are placed at the coordinates along the diffusion axes, which at a technical level are defined by the subdominant eigenvectors of the rate matrix describing the weak mutation dynamics [220] (see Methods for details), and squared distances have units of time measured in expected amino acid substitutions per site under neutrality. Boxes are drawn around the 3 main regions of functional sequences and frequency logos of those subsets of variants are drawn next to them.

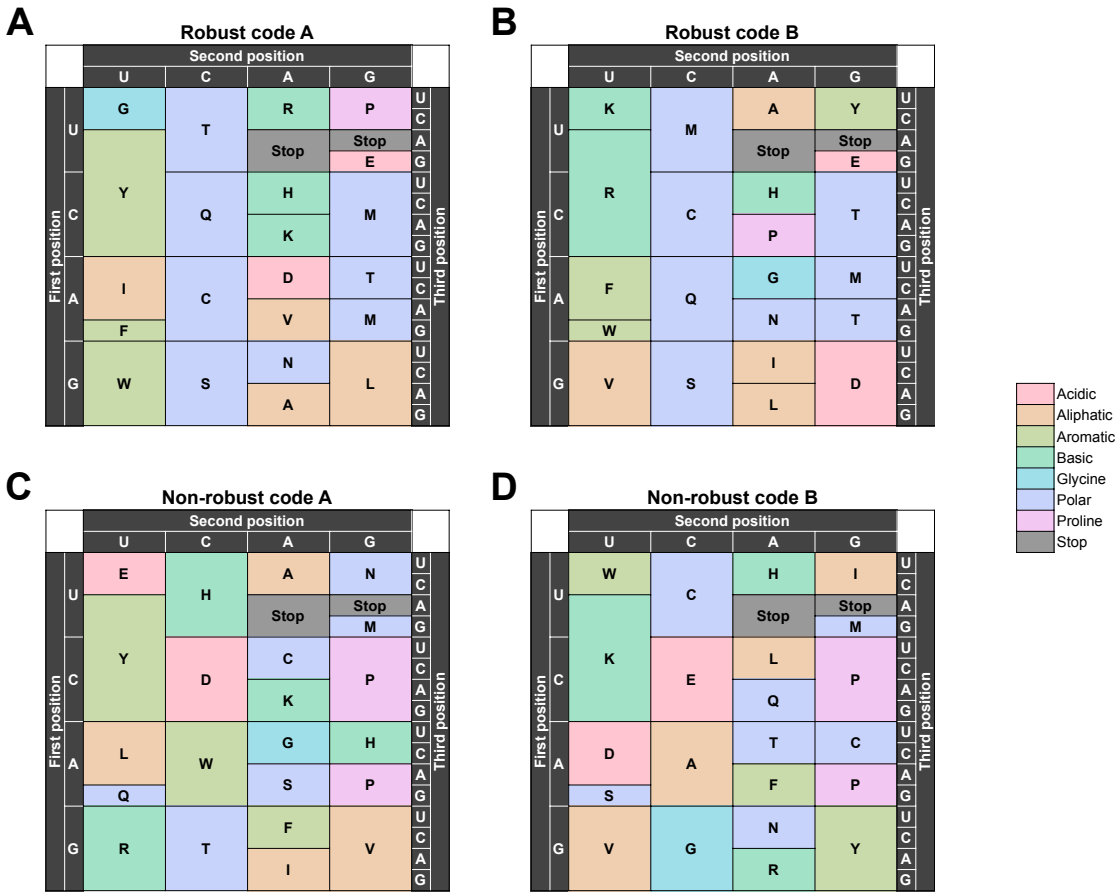


FIGURE S4.7: The amino acid permutation codes with (A, B) the highest and (C, D) lowest level of robustness. Codons are colored based on the physicochemical properties of the encoded amino acid, following Pines et al. [93].

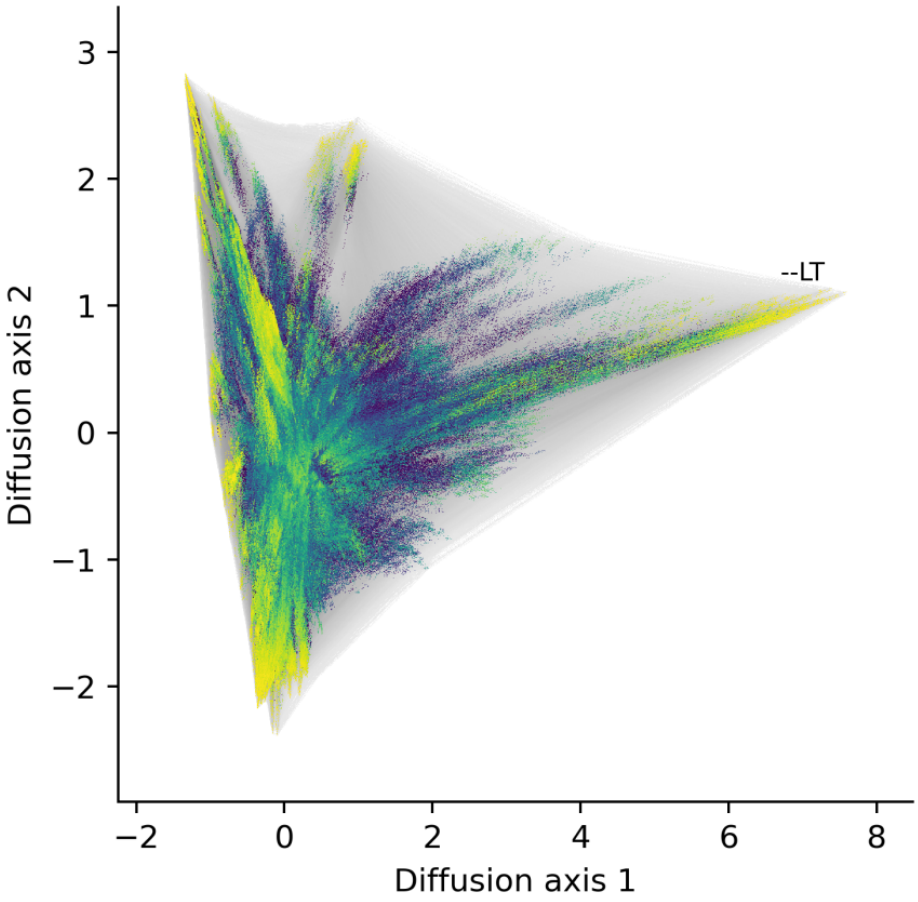


FIGURE S4.8: Visualization of the GB1 landscape under Robust Code B. The cluster of 41L-54T variants are indicated, which are separated from the rest of the landscape along Diffusion Axis 1. See Fig. 4.4 for further information on the layout, as well as the meaning of vertices and edges, and see Fig. 4.4C for an additional visualization of this landscape along Diffusion Axis 3.

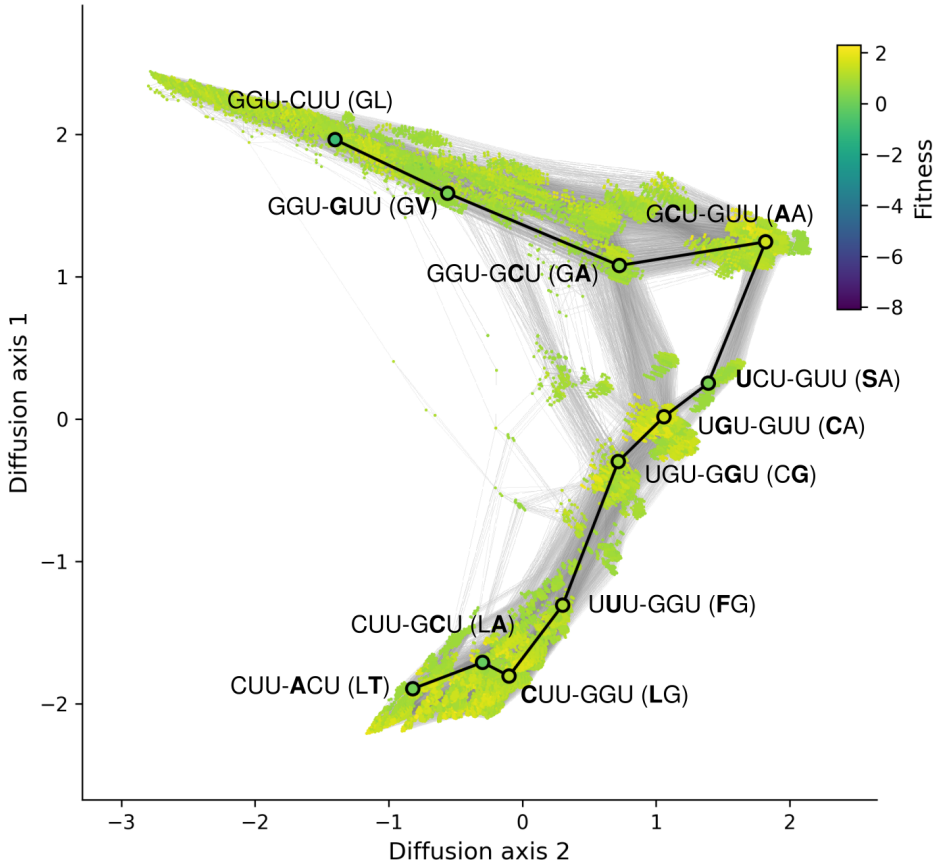


FIGURE S4.9: Visualization of the genotype network of high-fitness variants in the GB1 landscape, under the standard genetic code. The highlighted mutational path connects an mRNA sequence for 41G-54L to an mRNA sequence for 41L-54T via a series of intermediate mRNA sequences that are also part of the genotype network (i.e. that are among the 1% most fit sequences). Each highlighted edge corresponds to a non-synonymous point mutation, see Fig. 4.4 for further information on the layout. The modified nucleotide and amino acid are shown in bold at each step in the path. Color bar indicates protein fitness.

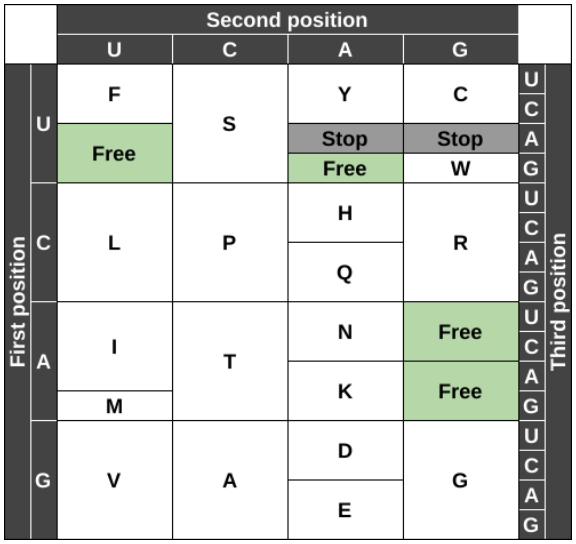


FIGURE S4.10: The codon table of the 57-codon *E. coli* genome [187], highlighting the four codon blocks that have been freed for reassignment.

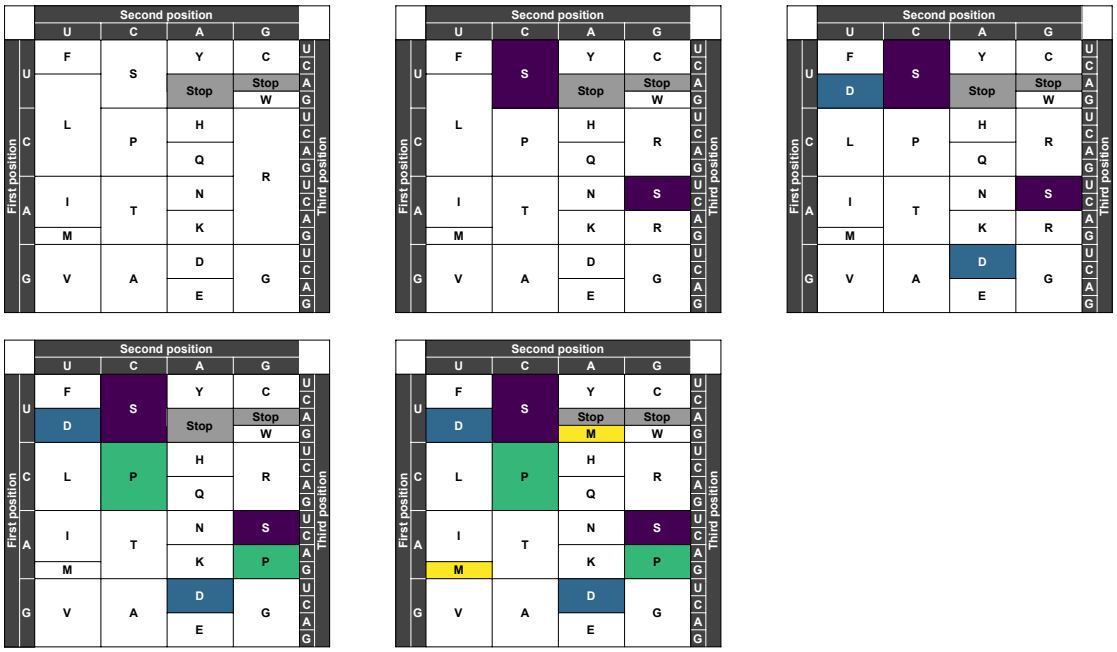


FIGURE S4.11: Examples of Ostrov codes with 0 to 4 split codon blocks, which are highlighted in colors.

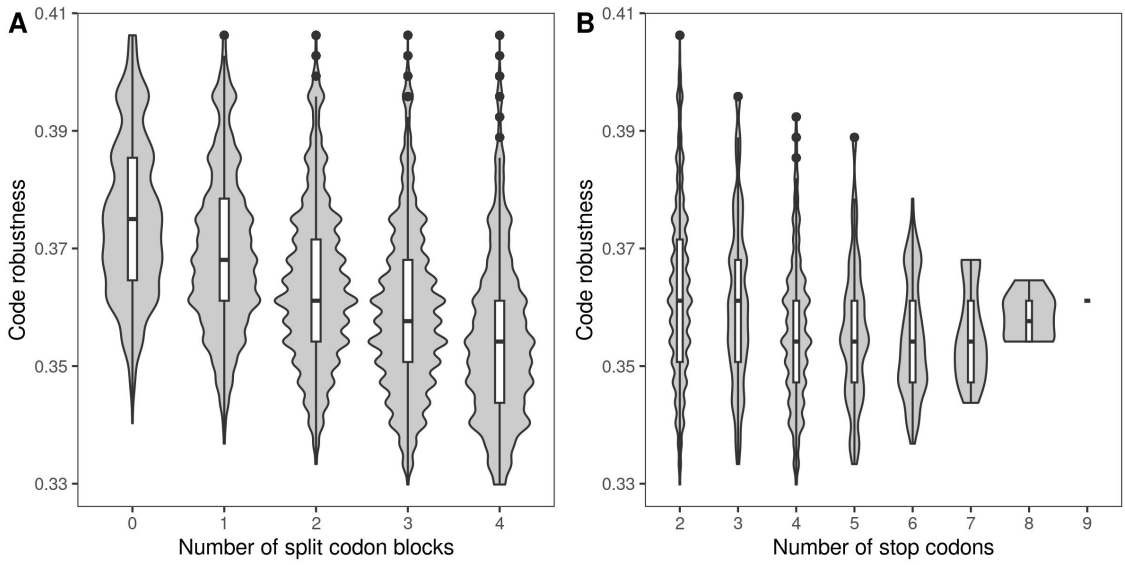


FIGURE S4.12: Violin plots of code robustness in relation to (A) the number of split codon blocks and (B) the number of stop codons in the 194,481 Ostrov codes. The violin plots show the distribution and the box-and-whisker plots the median, 25th and 75th percentile of code robustness (see Fig. S4.4 for details).

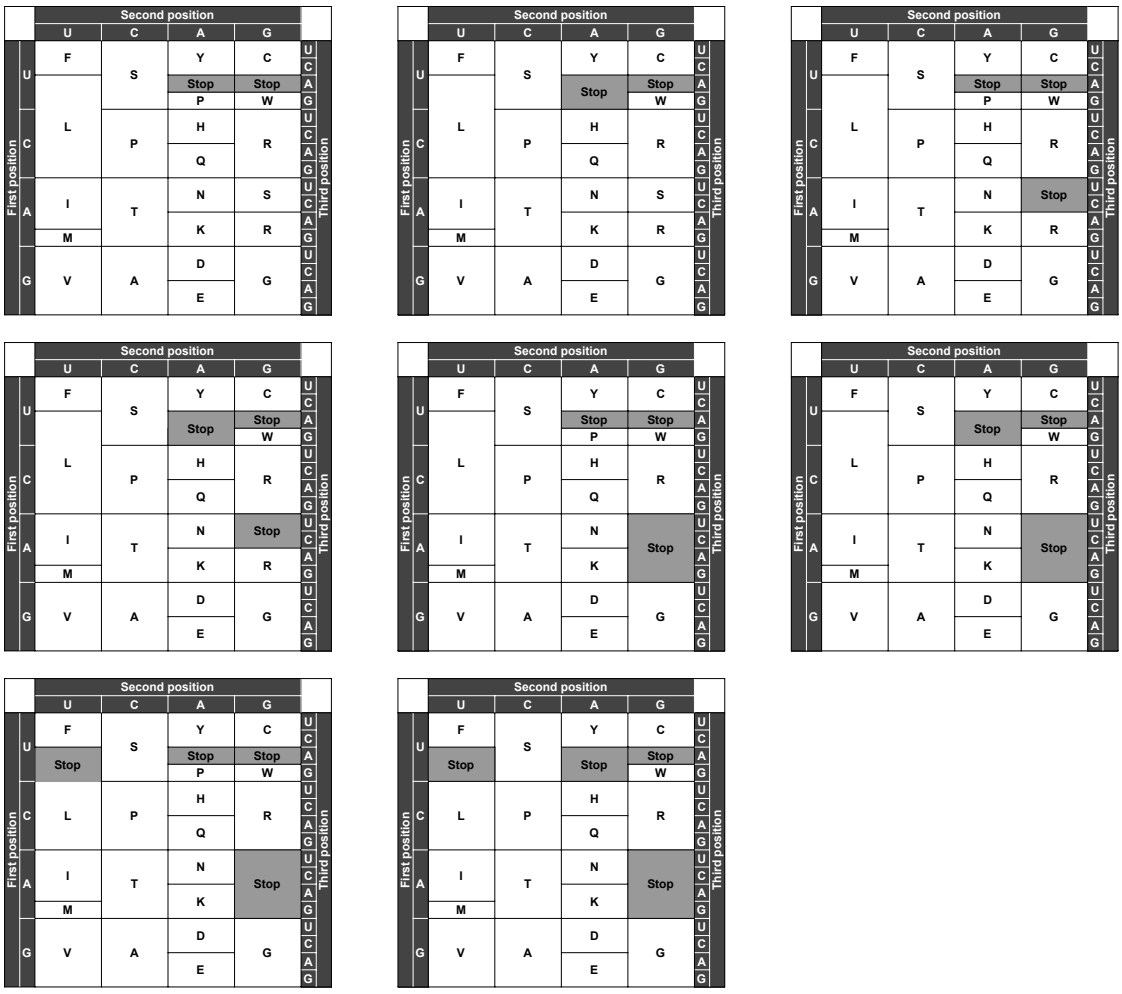


FIGURE S4.13: Examples of Ostrov codes with (top left) 2 to (bottom right) 9 stop codons, which are highlighted in grey.

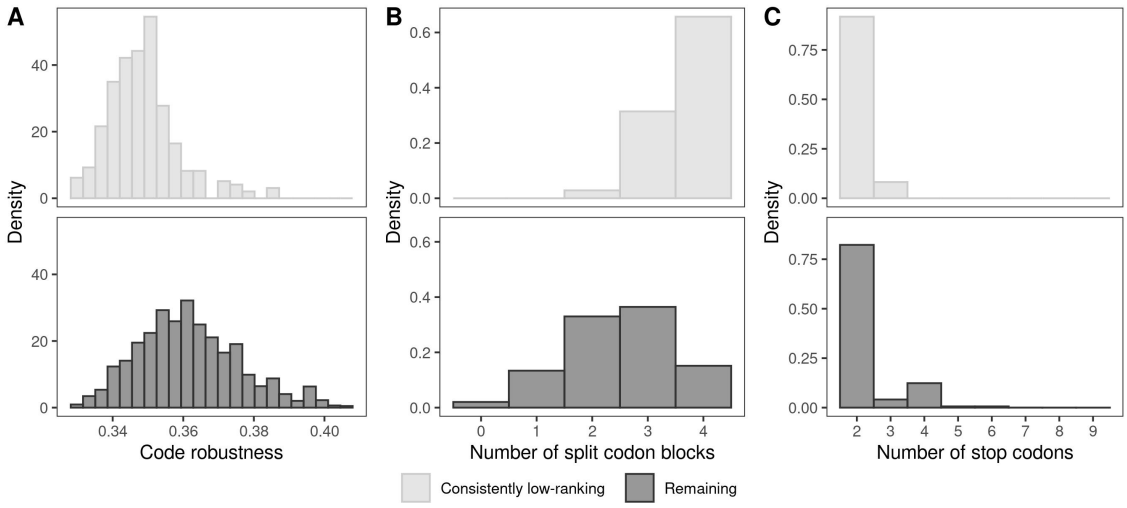


FIGURE S4.14: Comparison of the properties of the 280 consistently low-ranking codes with the remaining 194,201 codes, in terms of (A) code robustness, (B) number of split codon blocks, and (C) number of stop codons.

S4.2 *Supplementary tables*

	Proportion of reciprocal sign epistasis	Accessibility of the global peak
GB1		
Number of adaptive peaks	0.459	0.237
Proportion of reciprocal sign epistasis		0.160
ParD-ParE2		
Number of adaptive peaks	0.404	0.284
Proportion of reciprocal sign epistasis		0.384
ParD-ParE3		
Number of adaptive peaks	0.373	0.201
Proportion of reciprocal sign epistasis		0.057
ParB-*parS*		
Number of adaptive peaks	0.515	0.429
Proportion of reciprocal sign epistasis		0.511
ParB-*NBS*		
Number of adaptive peaks	0.624	0.496
Proportion of reciprocal sign epistasis		0.691
DHFR		
Number of adaptive peaks	0.229	0.241
Proportion of reciprocal sign epistasis		0.078

TABLE S4.1: Correlations of the individual measures of ruggedness with each other, for the six data sets.

	GB1	ParD-ParE2	ParD-ParE3	ParB-*parS*	ParB-*NBS*	DHFR
Number of peaks	-0.144	-0.119	-0.0346	-0.049	-0.103	-0.084
Prevalence of no epistasis	-	-	-	-	-	0.128
Prevalence of magnitude epistasis	0.247	0.219	0.118	0.097	0.148	-0.128; 0.102
Prevalence of simple sign epistasis	-0.125	-0.125	-0.0444	-0.020	-0.081	-0.104; -0.043
Prevalence of reciprocal sign epistasis	-0.277	-0.262	-0.191	-0.137	-0.209	-0.168; -0.200
Accessibility of the global peak	0.024	0.0052 $p = 0.099$	-0.151	0.097	0.035	0.078
Accessibility of the global peak (codes preserving the size of the global peak)	0.086	0.0920	-0.0846	0.065	0.097	0.111

TABLE S4.2: Correlation of various measures of landscape ruggedness with code robustness for amino acid permutation codes. All correlations are statistically significant, unless stated otherwise. Results for the prevalence of no epistasis are shown only for the DHFR landscape, because all amino acid permutation codes have the same proportion of squares with no epistasis in the remaining five landscapes: As the fitness values of all protein variants are distinct in these five ladscapes, only squares that involve at least two synonymous mutations exhibit no epistasis, and because all 100,000 amino acid permutation codes have the same block structure, the prevalence of such squares is the same for all of them. This is not true in the DHFR landscape, because of the non-functional variants, which are all assigned the same fitness value. For magnitude, simple sign, and reciprocal sign epistasis in the DHFR landscape, the first number is the correlation between the absolute prevalence of the corresponding type of epistasis and code robustness, while the second one is the correlation between code robustness and the prevalence of a given type of epistasis among epistatic squares only. The size of the subset of codes that preserve the size of the global peak and under which the global peak forms a single connected region in the genotype space (last row of the table) is $n = 3,769$, GB1; $n = 12,059$, ParD-ParE2; $n = 6,781$, ParD-ParE3; $n = 2,301$, ParB-*parS*; $n = 3,257$, ParB-*NBS*; $n = 6,852$, DHFR.

		GB1	ParD-ParE2	ParD-ParE3	ParB-*parS*	ParB-*NBS*	DHFR
Ruggedness	Number of peaks	0.00037	0.326	0.570	0.0229	0.046	0.205
	Prevalence of no epistasis	-	-	-	-	-	0.892
	Prevalence of magnitude epistasis	0.975	0.664	0.341	0.750	0.973	0.097; 0.240
	Prevalence of simple sign epistasis	0.123	0.548	0.604	0.481	0.103	0.323; 0.832
	Prevalence of reciprocal sign epistasis	0.0169	0.189	0.676	0.150	0.00404	0.154; 0.317
	Accessibility of the global peak	0.538; 0.864	0.609; 0.610	0.117; 0.643	0.688; 0.681	0.719; 0.590	0.622; 0.983
Greedy walks	Mean fitness	0.812; 0.857	0.961; 0.970	0.647; 0.864	0.953; 0.967	0.988; 0.975	0.186; 0.277
	Mean number of steps	0.936; 0.903	0.993; 0.995	0.928; 0.844	0.415; 0.282	0.995; 0.998	0.385; 0.980
	Entropy of the distribution of reached peaks	0.00304; 0.0024	0.00034; 0.00041	0.189; 0.084	0.0139; 0.0148	0.0024; 0.00675	0.490; 1.000
Weak mutation walks	Mean fitness after 500 mutations, $N = 10$	0.781; 0.839	0.756; 0.749	0.447; 0.807	0.975; 0.990	0.951; 0.943	0.071; 0.123
	Mean fitness after 500 mutations, $N = 100$	0.832; 0.884	0.795; 0.802	0.385; 0.748	0.960; 0.983	0.852; 0.738	0.171; 0.277
	Mean fitness after 500 mutations, $N = 10,000$	0.842; 0.893	0.797; 0.806	0.404; 0.775	0.961; 0.980	0.824; 0.662	0.257; 0.410
	Mean fitness after 500 mutations, $N = 1,000,000$	0.842; 0.894	0.796; 0.805	0.404; 0.775	0.961; 0.980	0.824; 0.660	0.255; 0.408

TABLE S4.3: Proportion of the amino acid permutation codes with lower or equal value of a given characteristic, as compared to the standard genetic code. Results for no epistasis shown only for DHFR, as it is the only landscape in which proportion of squares exhibiting no epistasis differs among amino acid permutation codes (see description of Supp. Tab. S4.2). For magnitude, simple sign, and reciprocal sign epistasis in the DHFR landscape, the first number gives the proportion for the absolute prevalence of the corresponding type of epistasis, while the second one for the prevalence of a given type of epistasis among epistatic squares only. For accessibility of the global peak and the greedy and weak mutation adaptive walks, two proportions are shown: the first one is the proportion of such codes in the whole data set of 100,000 amino acid permutation codes, the second is based on the subset of codes that preserve the size of the global peak and under which the global peak forms a single connected region in the genotype space ($n = 3,769$, GB1; $n = 12,059$, ParD-ParE2; $n = 6,781$, ParD-ParE3; $n = 2,301$, ParB-*parS*; $n = 3,257$, ParB-*NBS*; $n = 6,852$, DHFR). Note that in the ParB-*parS* data set, the standard genetic code is not a member of this set, as the global peak sequence in this landscape is RCWS and S is encoded by the split codon block in the standard genetic code.

	GB1	ParD-ParE2	ParD-ParE3	ParB-*parS*	ParB-*NBS*	DHFR
Mean fitness	0.107; 0.130	0.121; 0.183	-0.00422 ($p = 0.182$); 0.092	-0.0118; -0.0776	0.0731; 0.0468	-0.096; -0.082
Mean number of steps	0.192; 0.239	0.179; 0.170	0.157; 0.124	0.0454; 0.0654	0.0723; 0.0633	-0.002 ($p = 0.448$); -0.001 ($p = 0.918$)
Entropy of the distribution of reached peaks	-0.147; -0.203	-0.172; -0.213	-0.0702; -0.123	-0.0730; -0.0257 ($p = 0.218$)	-0.100; -0.0652	-0.065; -0.048

TABLE S4.4: Correlation of code robustness with the outcomes of greedy adaptive walks for amino acid permutation codes. The first number is the correlation in the set of 100,000 amino acid permutation codes, the second the correlation in the subset of codes that preserve the size of the global peak and under which the global peak is formed by a single connected region in the sequence space ($n = 3,769$, GB1; $n = 12,059$, ParD-ParE2; $n = 6,781$, ParD-ParE3; $n = 2,301$, ParB-*parS*; $n = 3,257$, ParB-*NBS*; $n = 6,852$, DHFR). All correlations are statistically significant, unless specified otherwise (p-values in parentheses).

Landscape	R	p-value
GB1	0.285	$p < 2.2 \cdot 10^{-16}$
ParD-ParE2	0.377	$p < 2.2 \cdot 10^{-16}$
ParD-ParE3	0.352	$p < 2.2 \cdot 10^{-16}$
ParB-*parS*	0.389	$p < 2.2 \cdot 10^{-16}$
ParB-*NBS*	0.614	$p < 2.2 \cdot 10^{-16}$
DHFR	0.219	$p < 2.2 \cdot 10^{-16}$

TABLE S4.5: Correlation between the size of the global peak and mean fitness reached by the greedy adaptive walks in the set of 100,000 amino acid permutation codes.

	GB1	ParD-ParE2	ParD-ParE3	ParB-*parS*	ParB-*NBS*	DHFR
Number of peaks	-0.412	-0.359	-0.300	-0.082	-0.145	-0.188
Prevalence of no epistasis	0.072	0.165	0.165	0.081	0.081	0.290
Prevalence of magnitude epistasis	-0.117; 0.0066, $p = 0.0035$	0.005, $p = 0.029$; 0.152	-0.132; 0.056	-0.265; -0.073	-0.141; -0.005 ($p = 0.036$)	-0.286; 0.193
Prevalence of simple sign epistasis	0.010; 0.087	-0.080; -0.010	-0.065; 0.005, $p = 0.018$	0.038; 0.140	0.013; 0.097	-0.298; 0.045
Prevalence of reciprocal sign epistasis	-0.180; -0.225	-0.309; -0.311	-0.198; -0.180	-0.102; -0.101	-0.176; -0.209	-0.357; -0.459
Accessibility of the global peak	0.142	-0.082	-0.230	0.234	0.197	0.150
Accessibility of the global peak (codes preserving the size of the global peak)	0.256	0.006, $p = 0.403$	0.039	0.248	0.507	0.387

TABLE S4.6: Correlation of various measures of landscape ruggedness with code robustness for the Ostrov codes. All correlations are statistically significant, unless stated otherwise. For magnitude, simple sign, and reciprocal sign epistasis, the first number is the correlation between the absolute prevalence of the corresponding type of epistasis and code robustness, while the second one is the correlation between code robustness and the prevalence of a given type of epistasis among epistatic squares only (i.e., in the second case, squares with no epistasis are discarded).

	GB1	ParD-ParE2	ParD-ParE3	ParB-*parS*	ParB-*NBS*	DHFR
Mean fitness	0.388; 0.357	0.208; 0.331	−0.076; 0.277	−0.134; −0.038	−0.018; 0.175	−0.313; −0.257
Mean number of steps	0.359; 0.289	0.266; 0.180	0.261; 0.170	−0.083; −0.018 ($p = 0.032$)	0.096; 0.271	−0.133; 0.064
Entropy of the distribution of reached peaks	−0.264; −0.183	−0.151; −0.220	0.040; −0.205	0.097; 0.032	−0.053; −0.151	−0.078; −0.214

TABLE S4.7: Correlation of code robustness with the outcomes of greedy adaptive walks for the Ostrov codes. The first number is the correlation in the complete set of 194,481 codes, the second the correlation in the subset of codes that preserve the size of the global peak and under which the global peak is formed by a single connected region in the sequence space. All correlations are statistically significant, unless stated otherwise.

	GB1	ParD-ParE2	ParD-ParE3	ParB-*parS*	ParB-*NBS*	DHFR
Number of peaks	0.847	0.487	0.657	0.655	0.631	0.627
Prevalence of no epistasis	−0.613	−0.699	−0.699	−0.624	−0.624	−0.225
Prevalence of magnitude epistasis	0.681; −0.380	0.313; −0.414	0.444; −0.364	0.658; −0.402	0.626; −0.392	0.217; −0.401
Prevalence of simple sign epistasis	0.452; 0.278	0.507; 0.322	0.485; 0.297	0.475; 0.310	0.473; 0.304	0.279; 0.144
Prevalence of reciprocal sign epistasis	0.614; 0.558	0.568; 0.444	0.576; 0.434	0.633; 0.588	0.618; 0.542	0.321; 0.528
Accessibility of the global peak	−0.094	−0.229	−0.256	−0.387	−0.078	−0.232

TABLE S4.8: Correlation of various measures of landscape ruggedness with number of split codon blocks for the Ostrov codes. All correlations are statistically significant. The mutational accessibility results pertain to the codes that preserve the size of the global peak and under which the global peak forms a single connected region in the genotype space. For magnitude, simple sign, and reciprocal sign epistasis, the first number is the correlation between the absolute prevalence of the corresponding type of epistasis and number of split codon blocks, while the second one is the correlation between number of split codon blocks and the prevalence of a given type of epistasis among epistatic squares only (i.e., in the second case, squares with no epistasis are discarded).

	GB1	ParD-ParE2	ParD-ParE3	ParB-*parS*	ParB-*NBS*	DHFR
Mean fitness	−0.261	−0.214	−0.109	−0.290	−0.347	0.063
Mean number of steps	−0.564	−0.348	−0.431	−0.339	−0.306	−0.081
Entropy of the distribution of reached peaks	0.453	0.360	0.316	0.289	0.174	0.146

TABLE S4.9: Correlation of number of split codon blocks with the outcomes of greedy adaptive walks for the Ostrov codes. All correlations are statistically significant.

	Number of split codon blocks	GB1	ParD-ParE2	ParD-ParE3	ParB-*parS*	ParB-*NBS*	DHFR
Number of peaks	0	0.460	0.213	0.359	0.066	0.072	0.224
	1	0.390	0.103	0.304	0.089	0.087	0.155
	2	0.330	0.073	0.245	0.110	0.104	0.127
	3	0.249	0.049	0.173	0.124	0.114	0.097
Prevalence of no epistasis	0	0.915	0.773	0.773	0.325	0.325	0.217
	1	0.922	0.789	0.789	0.291	0.291	0.154
	2	0.930	0.804	0.804	0.259	0.259	0.121
	3	0.938	0.822	0.822	0.233	0.233	0.084
Prevalence of magnitude epistasis	0	−0.444; 0.986	0.102; 0.906	0.061; 0.951	−0.174; 0.295	−0.213; 0.276	−0.199; 0.582
	1	−0.429; 0.981	0.154; 0.865	0.071; 0.893	−0.136; 0.270	−0.171; 0.254	−0.143; 0.535
	2	−0.395; 0.970	0.190; 0.806	0.088; 0.812	−0.100; 0.243	−0.132; 0.228	−0.112; 0.454
	3	−0.325; 0.939	0.189; 0.689	0.090; 0.664	−0.073; 0.216	−0.099; 0.198	−0.078; 0.333
Prevalence of simple sign epistasis	0	−0.983; −0.990	−0.959; −0.969	−0.953; −0.948	−0.330; −0.312	−0.321; −0.297	−0.489; −0.703
	1	−0.983; −0.987	−0.952; −0.953	−0.931; −0.905	−0.303; −0.292	−0.297; −0.281	−0.348; −0.596
	2	−0.982; −0.982	−0.940; −0.926	−0.893; −0.842	−0.276; −0.269	−0.272; −0.260	−0.267; −0.475
	3	−0.977; −0.966	−0.903; −0.860	−0.805; −0.713	−0.250; −0.244	−0.246; −0.236	−0.185; −0.331
Prevalence of reciprocal sign epistasis	0	−0.928; −0.883	−0.680; −0.495	−0.794; −0.688	−0.283; −0.223	−0.263; −0.184	−0.162; 0.087
	1	−0.918; −0.852	−0.626; −0.446	−0.753; −0.617	−0.244; −0.182	−0.225; −0.147	−0.116; 0.041
	2	−0.898; −0.802	−0.555; −0.386	−0.688; −0.533	−0.208; −0.145	−0.187; −0.111	−0.095; 0.015
	3	−0.843; −0.698	−0.438; −0.296	−0.568; −0.409	−0.175; −0.111	−0.149; −0.074	−0.069; −0.002 ($p = 0.547$)
Accessibility of the global peak	0	−0.488	−0.291	−0.457	−0.420	−0.327	−0.567
	1	−0.427	−0.262	−0.407	−0.448	−0.292	−0.365
	2	−0.330	−0.205	−0.341	−0.415	−0.211	−0.259
	3	-	-	−0.250	-	-	−0.174

TABLE S4.10: Correlation of various measures of landscape ruggedness with number of stop codons, conditioned on the number of split codon blocks, for the Ostrov codes. Results for 4 split codon blocks not shown because all codes with 4 split codon blocks have 2 stop codons. All correlations are statistically significant. The accessibility of the global peak results pertain to the codes that preserve the size of the global peak and under which the global peak forms a single connected region in the genotype space. For magnitude, simple sign, and reciprocal sign epistasis, the first number is the correlation between the absolute prevalence of the corresponding type of epistasis and number of stop codons, while the second one is the correlation between number of stop codons and the prevalence of a given type of epistasis among epistatic squares only (i.e., in the second case, squares with no epistasis are discarded). See Supp. Section S4.11 for the explanation of the epistasis results.

	Number of split codon blocks	GB1	ParD-ParE2	ParD-ParE3	ParB-*parS*	ParB-*NBS*	DHFR
Mean fitness	0	−0.364	−0.147	−0.183	$-6.85 \cdot 10^{-4}$ ($p = 0.966$)	−0.067	0.081
	1	−0.264	−0.128	−0.139	−0.033	−0.078	0.014 ($p = 0.025$)
	2	−0.196	−0.107	−0.104	−0.066	−0.095	−0.012
	3	−0.133	−0.078	−0.069	−0.090	−0.111	−0.022
Mean number of steps	0	−0.274	−0.079	−0.119	−0.041 ($p = 0.010$)	−0.064	0.005 ($p = 0.732$)
	1	−0.281	−0.074	−0.131	−0.055	−0.065	−0.036
	2	−0.277	−0.075	−0.135	−0.078	−0.074	−0.057
	3	−0.259	−0.074	−0.127	−0.103	−0.088	−0.054
Entropy of the distribution of reached peaks	0	0.310	0.064	0.182	0.019 ($p = 0.233$)	0.071	0.290
	1	0.295	0.068	0.181	0.045	0.066	0.252
	2	0.269	0.071	0.165	0.072	0.062	0.208
	3	0.222	0.064	0.132	0.093	0.060	0.144

TABLE S4.11: Correlation of the greedy adaptive walks outcomes with number of stop codons, conditioned on the number of split codon blocks, for the Ostrov codes. Results for 4 split codon blocks not shown because all codes with 4 split codon blocks have 2 stop codons. All correlations are statistically significant, unless stated otherwise.

	ParD-ParE2	ParD-ParE3	ParB-*parS*	ParB-*NBS*	DHFR
GB1	0.094	0.073	−0.022	−0.006	−0.277
ParD-ParE2		−0.045	0.007	0.021	−0.054
ParD-ParE3			−0.058	−0.013	−0.010
ParB-*parS*				0.059	−0.002
ParB-*NBS*					−0.008

TABLE S4.12: Correlation of the mean fitness reached in the greedy walks by individual Ostrov codes across the six data sets.

	GB1		ParD-ParE2		ParD-ParE3		ParB-*parS*		ParB-*NBS*		DHFR	
	Mean fitness	P-value	Mean fitness	P-value	Mean fitness	P-value	Mean fitness	P-value	Mean fitness	P-value	Mean fitness	P-value
Pines et al. [93]												
OPT	1.413	0.613	-0.063	0.498	-0.020	0.011	-2.727	0.774	-0.674	0.775	4.354	0.052
OPT-NR	1.207	0.949	-0.032	0.219	-0.045	0.056	-2.583	0.574	-1.652	1.0	4.155	0.177
CMC	1.372	0.722	0.0232	0.108	-0.0397	0.041	-3.160	1.0	-0.142	0.403	4.261	0.096
CMC^2	1.356	0.756	0.0495	0.059	-0.00585	0.004	-1.840	0.276	0.255	0.340	4.485	0.036
REC	1.449	0.503	-0.0143	0.153	-0.102	0.252	-2.496	0.488	-1.449	1.0	4.355	0.052
Ostrov	1.363	0.742	-0.0570	0.428	-0.0145	0.008	-2.674	0.679	-0.804	0.938	4.444	0.044
Calles et al. [195]												
FS20	-1.778	1.0	-2.864	1.0	-2.231	1.0	-7.061	1.0	-8.547	1.0	2.673	1.0
RED20	-0.871	1.0	-2.125	1.0	-0.939	1.0	-7.228	1.0	-8.288	1.0	3.361	1.0
This study												
standard	1.497	0.341	0.061	0.031	-0.125	0.350	-1.492	0.213	0.958	0.155	3.955	0.823
Code A	1.531	0.237	0.021	0.110	-0.057	0.093	-1.228	0.191	1.003	0.149	4.123	0.197
Code B	1.553	0.179	0.032	0.094	-0.066	0.128	-1.320	0.197	0.906	0.164	4.123	0.196
Code C	1.483	0.390	0.062	0.028	-0.046	0.058	-1.458	0.209	0.952	0.156	4.089	0.232
Code D	1.258	0.909	-0.142	0.955	-0.243	0.972	-3.156	1.0	-0.804	0.938	3.928	0.876

TABLE S4.13: Mean fitness reached in the greedy adaptive walks using the evolvability-promoting codes of Pines et al. [93], evolvability-diminishing codes of Calles et al. [195], and the codes identified in this study (codes A-C promote evolvability, code D diminishes). The p-value equals the proportion of the Ostrov codes that have reached the same or higher fitness in the greedy walks.

S4.3 *Artificial inflation of GB1 landscape ruggedness*

We reasoned that the weak correlation between code robustness and global peak accessibility might be due to the low variance in landscape ruggedness under the 100,000 amino acid permutation codes. To test this, we artificially inflated the ruggedness of the GB1 landscape under the standard genetic code by separately increasing the number of local peaks and the prevalence of reciprocal sign epistasis.

In particular, to increase the number of local peaks, we chose a number (ranging from 1 to 10,000) of protein sequences at random and set their fitness to a value that ensured that the corresponding region of the genotype network would form a local peak. In particular, we used a value of 2.5, which is halfway between the fitness value of the global peak (WWLA, fitness 2.52) and the second-best binding sequence (FYAA, fitness 2.48). Even when changing the fitness value of the same number of protein sequences, the number of local peaks in the resulting landscapes varies slightly, due to three reasons: (1) the original landscape contains 115 local peaks, some of which might cease to be local peaks if a neighboring sequence is artificially elevated; on the contrary, if a local peak is chosen and its elevation increased, the number of local peaks in the landscape does not change; (2) protein sequences containing the amino acid serine, which is encoded by the split codon block, are encoded by two disconnected regions in the genotype network, and artificially increasing their fitness thus creates two local peaks instead of one; (3) if two chosen sequences are neighbors in the genotype space, the corresponding mRNA sequences form one large plateau, and thus only one local peak is created instead of two.

To artificially increase the prevalence of reciprocal sign epistasis, we randomly sampled an mRNA sequence of length 12 (i.e., encoding 4 amino acids), making sure that it did not translate to the global peak sequence (WWLA) and that it did not contain any stop codons. We then sampled two mutations in different positions of the sequence that were non-synonymous, both alone and in combination, and such that the single mutants and double mutant did not contain a stop codon. We further required that the single mutants did not translate to the global peak sequence. We then permuted the fitness values of the corresponding protein sequences so that the double mutant had the highest value, the wild type the second highest, and the two single mutants the two lowest values. Since permuting the fitness values of a quadruplet of protein sequences changes the shape of many squares, it is again the case that even among landscapes in which the same number of squares was changed, the proportion of different types of epistasis is variable. We varied the number of squares forced to exhibit reciprocal sign epistasis from 1 to 100,000.

The resulting landscapes ranged in their number of local peaks from 115 to 3,356 and in their prevalence of reciprocal sign epistasis from 0.047 to 0.130. With these landscapes, we observed a strong correlation between global peak accessibility and the two measures of landscape ruggedness ($R = -0.893$, $p < 2.2 \cdot 10^{-16}$, number of peaks vs. mutational accessibility of the global peak; $R = -0.987$, $p < 2.2 \cdot 10^{-16}$, prevalence of reciprocal sign epistasis vs. mutational accessibility of the global peak; Supp. Fig. S4.15). Moreover, we observe that the moderate effect size is consistent with the range of reciprocal sign epistasis prevalence in the amino acid permutation codes (Supp. Fig. S4.15B) and that the expected effect size based on the range in the number of peaks would be even lower (Supp. Fig. S4.15A).

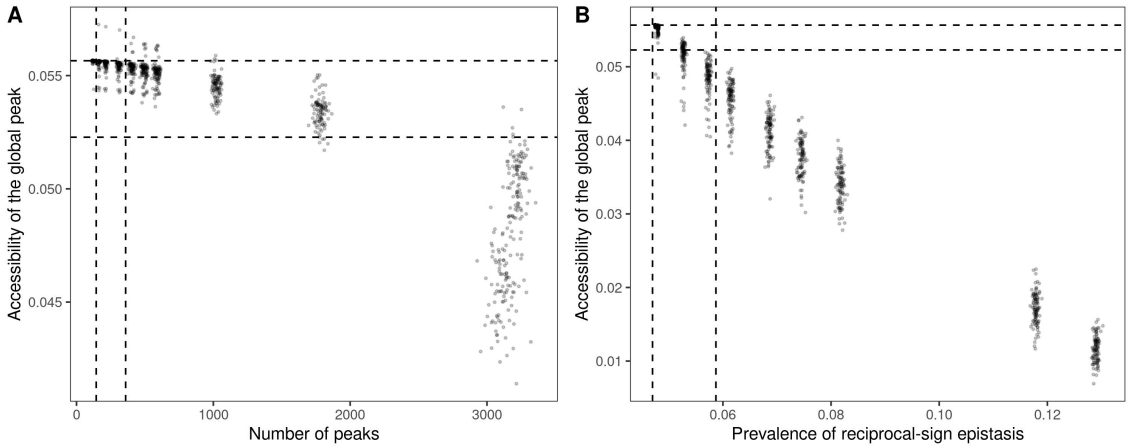


FIGURE S4.15: Accessibility of the global peak in relation to two artificially-inflated measures of landscape ruggedness, (A) the number of peaks and (B) the prevalence of reciprocal sign epistasis. The vertical lines show the 0.01 and 0.99 quantiles in the number of peaks and the prevalence of reciprocal sign epistasis, respectively, for the 100,000 amino acid permutation codes. The horizontal lines show the change in accessibility observed among the 100,000 amino acid permutation codes; the distance between them is the difference between the average accessibility of the global peak in landscapes generated using the 1% most and 1% least robust amino acid permutation codes. For visual clarity, the lines are positioned so that the top line coincides with the mutational accessibility of the global peak in the original landscape. Data pertain to the GB1 landscape under the standard genetic code.

S4.4 *Analysis of physicochemical properties from the AAindex database*

In the main text, we defined code robustness based on a discrete categorization of amino acids (Supp. Fig. S4.3). To see whether our conclusions also hold for other definitions of code robustness, and to discover which particular physicochemical properties of amino acids drive the correlation between code robustness and landscape ruggedness, here we additionally explore definitions of code robustness based on hundreds of different amino acid properties.

In particular, we downloaded the set of 566 different descriptors of amino acids from the AAindex database, version 9.2 [218], [219]. Of those, 13 contained at least one missing value; we discarded those, so the final set contained 553 different amino acid properties. These properties were previously divided into 4 categories: "alpha and turn propensity", "beta propensity", "hydrophobicity", and "other" [12]. For each property p and each of our 100,000 amino acid permutation codes we computed the mean absolute change in the property under genetic code c, $\mathrm{MAC}(p,c)$, as

$$\mathrm{MAC}(p,c) = \frac{1}{|\mathcal{V}|} \sum_{\{v,v'\}\in\mathcal{V}} |p[aa(v,c)] - p[aa(v',c)]|$$

where $\mathcal{V}$ is the set of all codon pairs that are one substitution away from each other (excluding pairs that contain at least one stop codon), $aa(v,c)$ is the amino acid encoded by codon v in genetic code c, and $p[a]$ is the value of property p for amino acid a. $\mathrm{MAC}(p,c)$ quantifies the sensitivity of code c with respect to amino acid property p: if $\mathrm{MAC}(p,c)$ is large, it means that single-nucleotide substitutions tend to cause large changes in property p; if $\mathrm{MAC}(p,c)$ is low, single-nucleotide substitutions tend to preserve property p.

For each property p we then computed Pearson's correlation between the sensitivities of our amino acid permutation codes with respect to property p and the different measures of landscape ruggedness:

$$R(p,\mathbf{T}) = \mathrm{corr}\left(\mathbf{MAC}(p,\mathbf{c}), \mathbf{T}(\mathbf{c})\right),$$

where $\mathbf{MAC}(p,\mathbf{c})$ is the vector of the values of $\mathrm{MAC}(p,c)$ for the set of 100,000 amino acid permutation codes and $\mathbf{T}(\mathbf{c})$ is the vector of values of certain landscape ruggedness measure (e.g., number of peaks) for these codes. For the mutational accessibility of the global peak, we only considered codes that preserve the size of the global peak, relative to the standard genetic code, and under which the global peak consists of a single connected region in the genotype space.

We determined the significance of each $R(p,\mathbf{T})$ by comparison with a null distribution calculated from 1,000,000 randomly generated amino acid "properties": We generated 1,000,000 null amino acid "properties" by uniformly sampling 20 random numbers between 0 and 1. For each such null property p_{null}, we computed $\mathbf{MAC}(p_{\mathrm{null}},\mathbf{c})$ and $R(p_{\mathrm{null}},\mathbf{T})$ as described above. The significance of the true correlation coefficient is then the proportion of these 1,000,000 null correlation coefficients that are more extreme than the true value. We then corrected the significance values for each landscape ruggedness measure for multiple testing using Benjamini-Hochberg correction [120].

The results are shown in Tab. S4.14. With the exception of the ParB-*NBS* landscape, we observe many amino acid properties that are consistent with our previous observation that more robust codes cause smoother adaptive landscapes, and only very few that support the opposite statement (i.e., less robust codes implying smoother landscapes). For example, for the GB1 data, 111 out

of the 553 properties (20.1% of the tested properties) exhibit a significant negative correlation between code robustness and the prevalence of reciprocal sign epistasis, whereas there are only 7 properties (i.e., 1.3% of the database) for which the opposite is true. For GB1, the statistically significant properties are enriched in beta-sheet propensity indices and, less consistently across the different landscape ruggedness measures, in alpha-helix propensity and hydrophobicity indices (Supp. Tab. S4.14). The importance of the preservation of hydrophobicity is consistent with three of the four residues (V39, G41, V54) being located in the protein core [260]; however, the consistent significance of beta-sheet propensity indices is somewhat surprising, as only one of the four residues is located in a beta sheet (Supp. Fig. S4.1). Similarly, the properties that are significant for the DHFR landscape are enriched in hydrophobicity indices, likely as a result of the fact that most peaks have D or E in their second position, which are the two most hydrophobic amino acids. The properties that are significant for the two ParD3 landscapes, as well as the ParB-*parS* landscape, in contrast, are consistently enriched in alpha-helix propensity indices (Supp. Tab. S4.14), which is consistent with both proteins being mostly helical (Supp. Fig. S4.1).

In sum, the amino acid properties most relevant to code robustness are protein-specific, and depend on the structural and functional properties of the assayed residues in each protein. Increasing code robustness relative to these properties generally results in smoother adaptive landscapes.

	Total	Alpha and turn propensity		Beta propensity		Hydrophobicity		Other	
	(+/- corr.)	Observed	P-value	Observed	P-value	Observed	P-value	Observed	P-value
GB1									
Number of peaks	171 (169/2)	14	1	23	**0.00937**	127	$\mathbf{< 2.2 \cdot 10^{-16}}$	7	1
Magnitude epistasis	45 (5/40)	21	**0.00251**	9	**0.00815**	12	0.973	3	1.00
Simple sign epistasis	0								
Reciprocal sign epistasis	118 (111/7)	50	$\mathbf{1.03 \cdot 10^{-4}}$	17	**0.0123**	43	0.762	8	1.00
Accessibility of the global peak	6 (5/1)	0	1	0	1	0	1	6	$\mathbf{3.53 \cdot 10^{-4}}$
ParD-ParE2									
Number of peaks	10 (5/5)	3	0.511	0	1	2	0.949	5	0.0980
Magnitude epistasis	62 (11/51)	34	$\mathbf{1.66 \cdot 10^{-6}}$	3	0.880	10	1	15	0.710
Simple sign epistasis	3 (0/3)	0	1	0	1	1	0.776	2	0.174
Reciprocal sign epistasis	89 (82/7)	56	$\mathbf{4.09 \cdot 10^{-13}}$	6	0.720	10	1	17	0.961
Accessibility of the global peak	1 (0/1)	1	0.262	0	1	0	1	0	1
ParD-ParE3									
Number of peaks	7 (2/5)	0	1	0	1	5	0.0891	2	0.594
Magnitude epistasis	72 (2/70)	56	$\mathbf{< 2.2 \cdot 10^{-16}}$	4	0.834	4	1	8	1
Simple sign epistasis	48 (44/4)	36	$\mathbf{2.41 \cdot 10^{-12}}$	4	0.537	3	1	5	0.998
Reciprocal sign epistasis	95 (95/0)	74	$\mathbf{< 2.2 \cdot 10^{-16}}$	5	0.882	4	1	12	1
Accessibility of the global peak	4 (4/0)	1	0.704	1	0.282	2	0.512	0	1
ParB-*parS*									
Number of peaks	13 (0/13)	0	1	0	1	10	**0.00664**	3	0.713
Magnitude epistasis	32 (2/30)	25	$\mathbf{1.29 \cdot 10^{-9}}$	3	0.474	3	1	1	1
Simple sign epistasis	38 (34/4)	28	$\mathbf{1.35 \cdot 10^{-9}}$	3	0.592	2	1	5	0.986
Reciprocal sign epistasis	30 (27/3)	22	$\mathbf{9.48 \cdot 10^{-8}}$	3	0.431	4	1	1	1
Accessibility of the global peak	0								

Continued on next page

Table S4.14 – continued from previous page

	Total	Alpha and turn propensity		Beta propensity		Hydrophobicity		Other	
	(+/- corr.)	Observed	P-value	Observed	P-value	Observed	P-value	Observed	P-value
ParB-*NBS*									
Number of peaks	13 (0/13)	0	1	0	1	12	$\mathbf{1.11 \cdot 10^{-4}}$	1	0.982
Magnitude epistasis	15 (15/0)	0	1	0	1	14	$\mathbf{1.95 \cdot 10^{-5}}$	1	0.990
Simple sign epistasis	21 (0/21)	0	1	0	1	20	$\mathbf{9.86 \cdot 10^{-8}}$	1	0.999
Reciprocal sign epistasis	10 (1/9)	0	1	1	0.564	9	**0.00143**	0	1
Accessibility of the global peak	2 (0/2)	0	1	0	1	2	0.154	0	1
DHFR									
Number of peaks	0								
No epistasis	20 (0/20)	0	1	2	0.480	18	$\mathbf{3.65 \cdot 10^{-6}}$	0	1
Magnitude epistasis	10 (0/10)	1	0.952	0	1	9	$\mathbf{1.43 \cdot 10^{-3}}$	0	1
Simple sign epistasis	5 (5/0)	1	0.781	0	1	4	0.081	0	1
Reciprocal sign epistasis	39 (39/0)	1	1	10	$\mathbf{7.50 \cdot 10^{-4}}$	27	$\mathbf{1.45 \cdot 10^{-4}}$	1	1
Accessibility of the global peak	14 (0/14)	0	1	1	0.687	13	$\mathbf{4.66 \cdot 10^{-5}}$	0	1

TABLE S4.14: Amino acid properties that are significantly correlated with landscape ruggedness. For each ruggedness measure, the total number of statistically significant properties is shown (number of positively/negatively correlated properties in parentheses). These properties are then broken down into four categories, for which the number of statistically significant properties is shown, along with the p-value of a one-tailed binomial test for over-abundance of the given category among the significantly correlated properties. Statistically significant p-values (< 0.05, **in bold**) mean that the category is over-represented among the significantly correlated properties, compared to the null expectation. The proportions of the categories in the database are 26.2% alpha and turn propensity, 8.0% beta propensity, 39.2% hydrophobicity, and 26.6% other. Results for no epistasis only shown for the DHFR landscape, because for the remaining landscapes the proportion of squares exhibiting no epistasis is the same under all amino acid permutation codes (see caption of Supp. Tab. S4.2). For DHFR, the correlations for magnitude, simple sign, and reciprocal sign epistasis were computed using the proportions of a given epistasis type among epistatic squares only.

S4.5 *Data set-specific definition of code robustness*

In the main text, we showed that code robustness, defined as the proportion of "conservative" substitutions allowed by a code, exhibits a weak inverse correlation with landscape ruggedness. To define "conservative" substitutions, we categorized amino acids into discrete groups (Supp. Fig. S4.3), based on the long-standing observation that some amino acid pairs are more exchangeable than others [235]–[237]. However, because it was only possible to perform our analyses on a small number of landscapes, which were, in turn, based on only a small number of protein sites, the extent to which this measure of conservativeness applies to each of our data sets is not immediately apparent. For instance, it may be the case that the amino acid exchangeabilities in our individual data sets are markedly different from the general principles of exchangeability deduced from a large number of proteins and protein sites.

To test whether this is the case, we considered a data set-specific definition of code robustness, calculated as the expected change in fitness upon mutation. To do so, we first calculated an empirical amino acid exchangeability matrix for each data set, in which each matrix entry reports the mean absolute change in fitness for the corresponding pair of amino acids across all possible genetic backgrounds (Supp. Fig. S4.16). For the DHFR landscape, we only considered functional variants. We then defined code robustness as the mean amino acid exchangeability across all possible single-nucleotide substitutions, excluding mutations to or from stop codons.

For each data set, we correlated the empirical amino acid exchangeabilities with those based on physicochemical properties, for all 553 amino acid properties from the AAindex database, separately [218], [219] (see also Supp. Section S4.4). The distributions of correlation coefficients are shown in Supp. Fig. S4.17 and the most strongly correlated properties for each data set are listed in Supp. Tab. S4.15. Whereas the empirical amino acid exchangeabilities exhibit strong correlations with physicochemical properties in some cases (e.g., for DHFR, the empirical exhangeabilities exhibit a correlation of $R = 0.983$ with the property FAUJ880112, which lists the negative charge of amino acids), in others they exhibit at best only moderate correlations (e.g., for ParD-ParE2, the strongest correlation is $R = 0.503$, specifically with the property SNEP660101, a measure of alpha-helix propensity). Moreover, for ParB-*NBS*, the empirical exchangeabilities exhibit a negative correlation with the majority of the 553 physicochemical properties (Supp. Fig. S4.17E). There is therefore often a disconnect between measures of amino acid exchangeability based on the physicochemical properties of amino acids and those based on fitness changes observed in our individual data sets. This helps explain the weak correlations between code robustness and landscape ruggedness reported in the main text, as well as the variation in the strength of this correlation across data sets. Indeed, when we define code robustness based on the empirical exchangeability matrices, using 100,000 codes generated by amino acid permutation restricted to preserve the number of codons per amino acid (see Supp. Section S4.6), we consistently observe significant inverse correlations with landscape ruggedness for all six data sets (Supp. Tab. S4.16). This observation lends additional support to the hypothesis that robust genetic codes, i.e. codes that preferentially allow conservative mutations, produce smooth fitness landscapes that enhance protein evolvability.

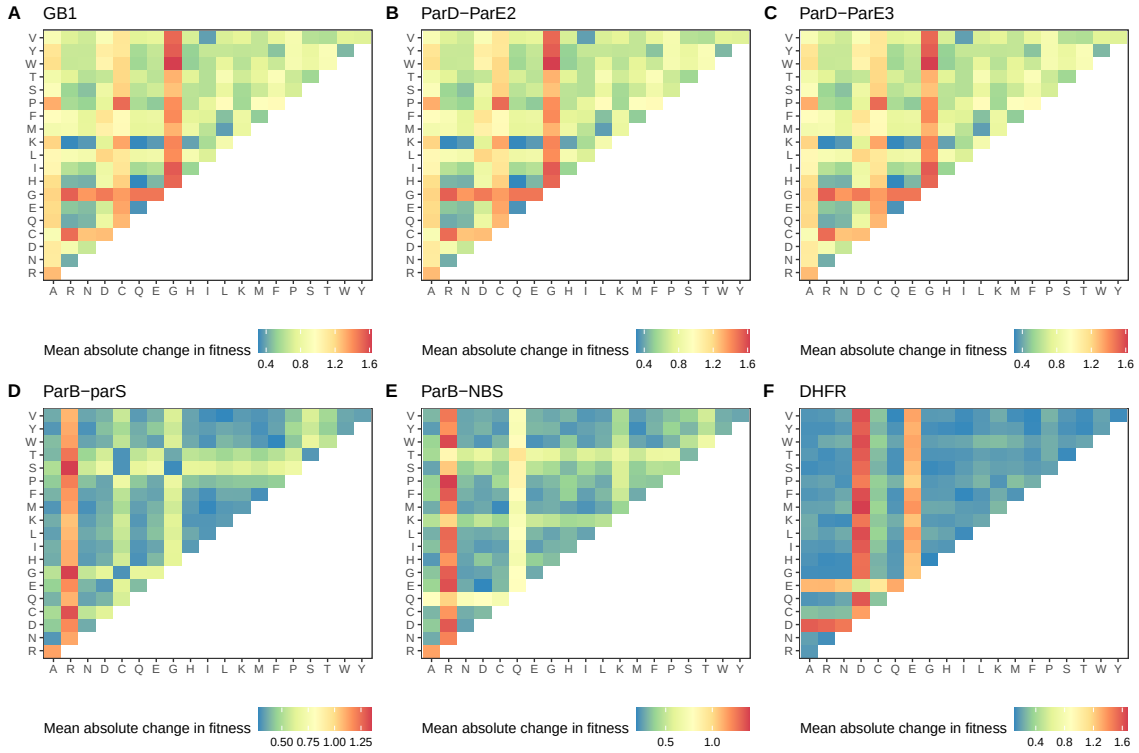


FIGURE S4.16: Empirical exchangeability matrices for (A) GB1, (B) ParD-ParE2, (C) ParD-ParE3, (D) ParB-parS, (E) ParB-NBS, and (F) DHFR. Each matrix element shows the mean absolute change in fitness across all genetic backgrounds. For DHFR, only functional sequences were used for the computation.

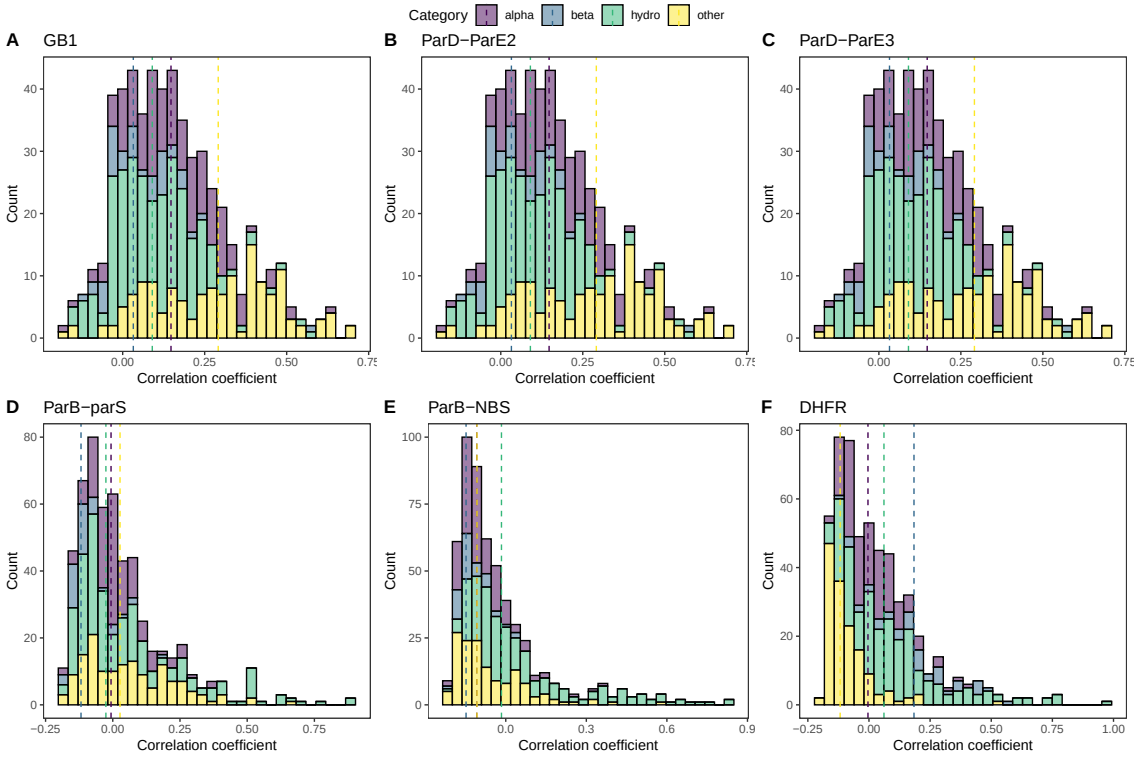


FIGURE S4.17: Distribution of correlation coefficients for amino acid exchangeabilities measured in terms of absolute fitness change vs. absolute change in physicochemical property, for each of the 553 amino acid properties in the AAindex database [218], [219]. The properties are colored by their category, according to ref. [12]. The vertical lines denote the median correlation coefficient for each category.

Property	Description	Category	R	Reference
GB1				
KARS160109	Maximum eigenvalue of the weighted Laplacian matrix of the graph	other	0.702	[261]
KARS160108	Average weighted degree	other	0.682	[261]
DAWD720101	Size	other	0.649	[262]
LEVM760104	Side chain torsion angle phi(AAAR)	other	0.646	[263]
KARS160121	Weighted average eigenvalue based on the atomic numbers	other	0.625	[261]
KARS160111	Average eigenvalue of the Laplacian matrix of the the graph	other	0.621	[261]
KOEP990101	Alpha-helix propensity derived from designed sequences	alpha	0.620	[264]
KARS160119	Weighted maximum eigenvalue based on the atomic numbers	other	0.613	[261]
LEVM760103	Side chain angle theta(AAR)	other	0.613	[263]
MAXF760105	Normalized frequency of zeta L	other	0.594	[265]
ParD-ParE2				
SNEP660101	Principal component I	alpha	0.503	[266]
YUTK870101	Unfolding Gibbs energy in water, pH7.0	hydro	0.463	[267]
QIAN880135	Weights for coil at the window position of 2	alpha	0.454	[268]
COHE430101	Partial specific volume	other	0.452	[269]
LEVM780106	Normalized frequency of reverse turn, unweighted	alpha	0.451	[270]
TANS770110	Normalized frequency of chain reversal	alpha	0.450	[271]
NAKH900112	Transmembrane regions of mt-proteins	other	0.446	[272]
FAUJ880102	Smoothed upsilon steric parameter	other	0.440	[273]
NAKH920105	AA composition of MEM of single-spanning proteins	other	0.437	[274]
QIAN880133	Weights for coil at the window position of 0	alpha	0.436	[268]
ParD-ParE3				
FINA910104	Helix termination parameter at posision j+1	alpha	0.698	[275]
MUNV940105	Free energy in beta-strand region	beta	0.639	[276]
ROBB760104	Information measure for C-terminal helix	alpha	0.630	[277]
QIAN880109	Weights for alpha-helix at the window position of 2	alpha	0.616	[268]
FAUJ880113	pK-a(RCOOH)	alpha	0.606	[273]
MUNV940101	Free energy in alpha-helical conformation	alpha	0.605	[276]
MUNV940104	Free energy in beta-strand region	beta	0.588	[276]
BLAM930101	Alpha helix propensity of position 44 in T4 lysozyme	alpha	0.587	[278]
FINA910102	Helix initiation parameter at posision i,i+1,i+2	alpha	0.586	[275]
QIAN880134	Weights for coil at the window position of 1	alpha	0.580	[268]

Continued on next page

Table S4.15 – continued from previous page

Property	Description	Category	R	Reference
ParB-*parS*				
YUTK870103	Activation Gibbs energy of unfolding, pH7.0	hydro	0.876	[267]
YUTK870104	Activation Gibbs energy of unfolding, pH9.0	hydro	0.875	[267]
EISD860102	Atom-based hydrophobic moment	hydro	0.771	[279]
JACR890101	Weights from the IFH scale	hydro	0.684	[280]
JOND750102	pK (-COOH)	hydro	0.679	[281]
HUTJ700103	Entropy of formation	other	0.652	[282]
YUTK870102	Unfolding Gibbs energy in water, pH9.0	hydro	0.634	[267]
FAUJ880109	Number of hydrogen bond donors	hydro	0.621	[273]
KHAG800101	The Kerr-constant increments	hydro	0.621	[283]
RADA880104	Transfer free energy from chx to oct	hydro	0.564	[284]
ParB-*NBS*				
YUTK870103	Activation Gibbs energy of unfolding, pH7.0	hydro	0.822	[267]
YUTK870104	Activation Gibbs energy of unfolding, pH9.0	hydro	0.820	[267]
EISD860102	Atom-based hydrophobic moment	hydro	0.755	[279]
FAUJ880109	Number of hydrogen bond donors	hydro	0.727	[273]
JACR890101	Weights from the IFH scale	hydro	0.679	[280]
RADA880107	Energy transfer from out to in(95%buried)	hydro	0.664	[284]
GUYH850105	Apparent partition energies calculated from Chothia index	hydro	0.655	[285]
RADA880104	Transfer free energy from chx to oct	hydro	0.623	[284]
YUTK870102	Unfolding Gibbs energy in water, pH9.0	hydro	0.573	[267]
JOND750102	pK (-COOH)	hydro	0.572	[281]
DHFR				
FAUJ880112	Negative charge	other	0.983	[273]
RICJ880105	Relative preference value at N2	hydro	0.790	[286]
RICJ880106	Relative preference value at N3	hydro	0.788	[286]
HOPA770101	Hydration number	hydro	0.781	[287]
WOEC730101	Polar requirement	hydro	0.715	[288]
FINA910101	Helix initiation parameter at posision i-1	hydro	0.710	[275]
CHOP780204	Normalized frequency of N-terminal helix	hydro	0.667	[289]
KLEP840101	Net charge	hydro	0.656	[290]
ROBB760102	Information measure for N-terminal helix	hydro	0.628	[277]
AURR980107	Normalized positional residue frequency at helix termini N2	hydro	0.601	[291]

TABLE S4.15: Top 10 properties most strongly correlated with the empirical amino acid exchangeabilities, for the six data sets. Property codes and descriptions as in the AAindex database. Categorization according to ref. [12].

	GB1	ParD-ParE2	ParD-ParE3	ParB-*parS*	ParB-*NBS*	DHFR
Number of peaks	0.416	0.052	0.213	0.049	0.133	0.319
Prevalence of no epistasis	-	-	-	-	-	−0.750
Prevalence of magnitude epistasis	−0.295	−0.217	−0.056	0.248	−0.386	0.586
Prevalence of simple sign epistasis	−0.267	−0.161	−0.397	−0.485	0.033	−0.676
Prevalence of reciprocal sign epistasis	0.572	0.567	0.682	0.220	0.557	0.030
Accessibility of the global peak	−0.093	−0.369	−0.017	−0.192	−0.457	−0.486

TABLE S4.16: Correlation of various measures of landscape ruggedness with an alternative definition of code robustness that is based on the mean fitness change upon mutation, as measured in a specific data set. All correlations are statistically significant. Results for the prevalence of no epistasis are shown only for DHFR, because in the remaining landscapes all amino acid permutation codes have the same proportion of squares with no epistasis. For DHFR, the magnitude, simple sign, and reciprocal sign epistasis results correspond to correlation between code robustness and the proportion of given epistasis type among epistatic squares only.

S4.6 *Restricted amino acid permutation codes*

In the standard genetic code, the number of codons encoding an amino acid ranges from 1 (M, W) to 6 (L, S, R). Because the classical amino acid permutation rewiring scheme does not impose any restrictions on the permutations, the number of codons encoding amino acid *X* differs amongst rewired codes. This is important from the point of view of code robustness, as code robustness increases if "special" amino acids, such as glycine or proline, are assigned to small codon blocks. On the other hand, it presents a challenge for the interpretation of our results, because the number of mRNA sequences encoding a given protein variant differs amongst codes, which means different codes yield landscapes with different mean fitness. Moreover, as shown in the main text, the number of mRNA sequences encoding the global peak influences its mutational accessibility and the outcomes of greedy walks.

To check that our results also hold for genetic codes that preserve the number of codons per amino acid, we generated 100,000 genetic codes by amino acid permutation, but restricted the permutations to those that do not change the number of codons encoding each amino acid relative to the standard genetic code. Under these genetic codes, the number of mRNAs encoding each protein variant is exactly the same as in the standard genetic code, and the resulting fitness landscapes thus differ only by which protein variants are reachable from each other and which are not. We then repeated the analyses from the main text using these genetic codes.

We observe qualitatively the same results. In particular, we again observe that more robust codes result in smoother fitness landscapes (Supp. Tab. S4.17). Consequently, we also see that more robust codes lead to higher fitness reached in simulations of evolution (with the exception of the DHFR landscape) and increased predictability of evolution (Supp. Tab. S4.18).

	GB1	ParD-ParE2	ParD-ParE3	ParB-*parS*	ParB-*NBS*	DHFR
Number of peaks	−0.115	−0.096	−0.057	−0.109	−0.059	−0.003 ($p = 0.342$)
Prevalence of no epistasis	-	-	-	-	-	0.211
Prevalence of magnitude epistasis	0.269	0.055	−0.012	−0.012	0.186	−0.190; −0.102
Prevalence of simple sign epistasis	−0.090	0.075	0.156	0.171	−0.107	0.115; 0.142
Prevalence of reciprocal sign epistasis	−0.281	−0.188	−0.214	−0.188	−0.179	−0.127; −0.066
Accessibility of the global peak	0.109	0.171	0.073	0.121	0.017	0.130

TABLE S4.17: Correlation of various measures of landscape ruggedness with code robustness, for amino acid permutation codes restricted to preserve the number of codons per amino acid. All correlations are statistically significant, unless stated otherwise. Results for the prevalence of no epistasis are shown only for the DHFR landscape, because in the remaining 5 landscapes all amino acid permutation codes have the same proportion of squares with no epistasis. For DHFR, two numbers are shown for the prevalence of magnitude, simple sign, and reciprocal sign epistasis; the first one is the correlation between code robustness and the prevalence of a given type of epistasis among all squares, the second is the correlation between code robustness and the prevalence of a given type of epistasis among epistatic squares only. The results for accessibility of the global peak are based on a subset of the genetic codes under which the global peak is formed by single connected region in the sequence space.

	GB1	ParD-ParE2	ParD-ParE3	ParB-*parS*	ParB-*NBS*	DHFR
Mean fitness	0.066	0.229	0.077	$-6.24 \cdot 10^{-4}$ ($p = 0.910$)	0.022	−0.071
Mean number of steps	0.078	0.220	0.082	−0.047	0.095	−0.068
Entropy of the distribution of reached peaks	−0.124	−0.275	−0.122	−0.078	−0.097	−0.004 ($p = 0.239$)

TABLE S4.18: Correlation of code robustness with the outcomes of greedy adaptive walks for amino acid permutation codes restricted to preserve the number of codons per amino acid as in the standard genetic code. The results are based on a subset of the genetic codes under which the global peak is formed by a single connected region in the sequence space. All correlations are statistically significant, unless stated otherwise.

S4.7 *Random codon assignment codes*

To further verify that our results are insensitive to the choice of rewiring scheme, we also repeated the analyses from the main text for 100,000 genetic codes generated by randomly assigning an amino acid meaning to each of the 61 sense codons, ensuring that each of the 20 amino acids is assigned at least one codon [95], [115], [121]. We refer to these as "random codon assignment" codes. These differ from the codes generated using amino acid permutation by lacking the synonymous codon block structure of the standard genetic code, making them less realistic as possible alternatives to the standard genetic code than the amino acid permutation codes. As a result of the missing block structure, their average robustness is much lower than that of the amino acid permutation codes ($p < 2.2 \cdot 10^{-16}$, Welch two sample t-test; Supp. Fig. S4.18).

Consistent with our previous observations, we find that increasing code robustness decreases landscape ruggedness under this alternative rewiring scheme (Supp. Tab. S4.19) and that code robustness is positively correlated with length of greedy adaptive walks and predictability of evolution (Supp. Tab. S4.20).

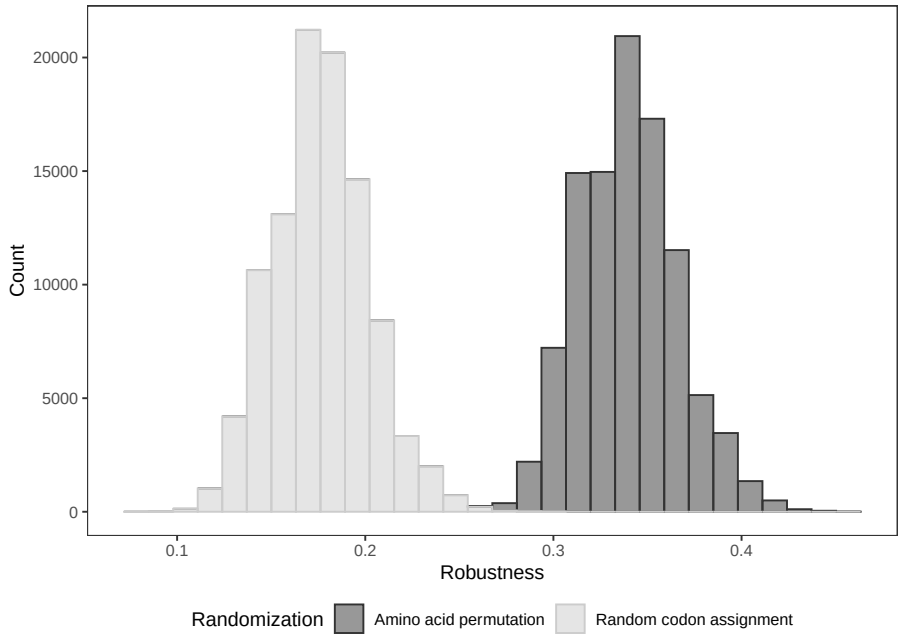


FIGURE S4.18: Robustness distributions for genetic codes rewired by amino acid permutation and random codon assignment. Data pertain to 100,000 codes per rewiring scheme.

	GB1	ParD-ParE2	ParD-ParE3	ParB-*parS*	ParB-*NBS*	DHFR
Number of peaks	−0.387	−0.392	−0.329	−0.296	−0.319	−0.283
Prevalence of no epistasis	0.371	0.371	0.371	0.371	0.371	0.189
Prevalence of magnitude epistasis	−0.160	0.133	−0.038	−0.273	−0.127	−0.185
Prevalence of simple sign epistasis	−0.417	−0.393	−0.249	−0.371	−0.362	−0.182
Prevalence of reciprocal sign epistasis	−0.476	−0.466	−0.401	−0.406	−0.447	−0.248
Accessibility of the global peak	−0.002, $p = 0.535$	−0.053	−0.228	0.123	0.022	0.047
Accessibility of the global peak (codes preserving the size of the global peak)	0.149	0.134	0.024, $p = 0.190$	0.018, $p = 0.149$	0.065	0.150

TABLE S4.19: Correlation of various measures of landscape ruggedness with code robustness for random codon assignment codes. All correlations are statistically significant, unless stated otherwise. Unlike for the amino acid permutation codes, the random codon assignment codes differ in the proportion of mutations that are synonymous, and hence also differ in the prevalence of squares showing no epistasis. However, the observed correlation with code robustness is exactly the same for all data sets except DHFR because the prevalence of squares showing no epistasis is determined by the proportion of synonymous mutations in the genetic code. (This is not true for DHFR because of the non-functional variants.)

	GB1	ParE2	ParE3	ParB-*parS*	ParB-*NBS*	DHFR
Mean fitness	0.140; 0.141	0.127; 0.183	−0.0734; 0.092	−0.021; −0.065	0.0869; 0.138	−0.130; −0.157
Mean number of steps	0.418; 0.454	0.259; 0.170	0.266; 0.124	0.164; 0.158	0.204; 0.210	0.114; 0.091
Entropy of the distribution of reached peaks	−0.071; −0.109	−0.072; −0.213	0.0635; −0.123	−0.084; −0.070	−0.0928; −0.118	−0.131; −0.129

TABLE S4.20: Correlation of code robustness with the outcomes of greedy adaptive walks for the random codon assignment codes. In each cell, the first number gives the correlation in the full set of 100,000 random codon assignment codes, the second number the correlation for the subset of genetic codes that preserve the size of the global peak. (For these codes, the subset of codes that preserve the size of the global peak is not required to also fulfill the condition that the global peak forms a single connected region in the genotype space, because due to the nature of the random codon assignment codes this is almost never the case.) The size of the subset is $n = 4,584$ (GB1), $n = 16,032$ (ParD-ParE2), $n = 6,781$ (ParD-ParE3), $n = 6,484$ (ParB-*parS*), $n = 3,099$ (ParB-*NBS*), $n = 3,059$ (DHFR). All correlations are statistically significant.

S4.8 *Landscape dimensionality*

One of the most fundamental characteristics of a fitness landscape is its dimensionality, which in our empirical landscapes is determined by the number of variable protein sites, L. Recent theoretical work has shown that landscapes of low dimensionality tend to be more rugged than their higher-dimensional counterparts [234]. We therefore wanted to test the sensitivity of our results, which all pertain to small L, to changes in landscape dimensionality. However, because the number of protein variants grows exponentially with L, it is currently experimentally intractable to obtain combinatorially-complete fitness landscapes of higher dimensionality.

We therefore studied the sensitivity of our results to landscape dimensionality by generating landscapes of even lower dimensionality. We did so by subsetting the GB1 data, which has the most variable sites of any of our data sets ($L = 4$), to contain only $L = 2$ or 3 variable sites. For $L = 3$, there are 80 such landscapes (4 ways to choose 3 sites out of 4×20 different backgrounds) and we generated all of them; for $L = 2$, there are 2,400 such landscapes (6 ways to choose 2 sites $\times 20^2$ different backgrounds) and we randomly sampled 100 of them. We then repeated our analyses from the main text on these low-dimensionality landscapes. The results are summarized in Supp. Fig. S4.19 and S4.20, which depict histograms of correlation coefficients between code robustness and our various measures of protein evolvability. While the variance in the distribution of correlation coefficients is high, code robustness is on average negatively correlated with landscape ruggedness and positively correlated with the mean fitness reached by greedy adaptive walks. Moreover, the strength of these trends tends to increase with L. Taken together, these results suggests our main conclusions are qualitatively insensitive to landscape dimensionality.

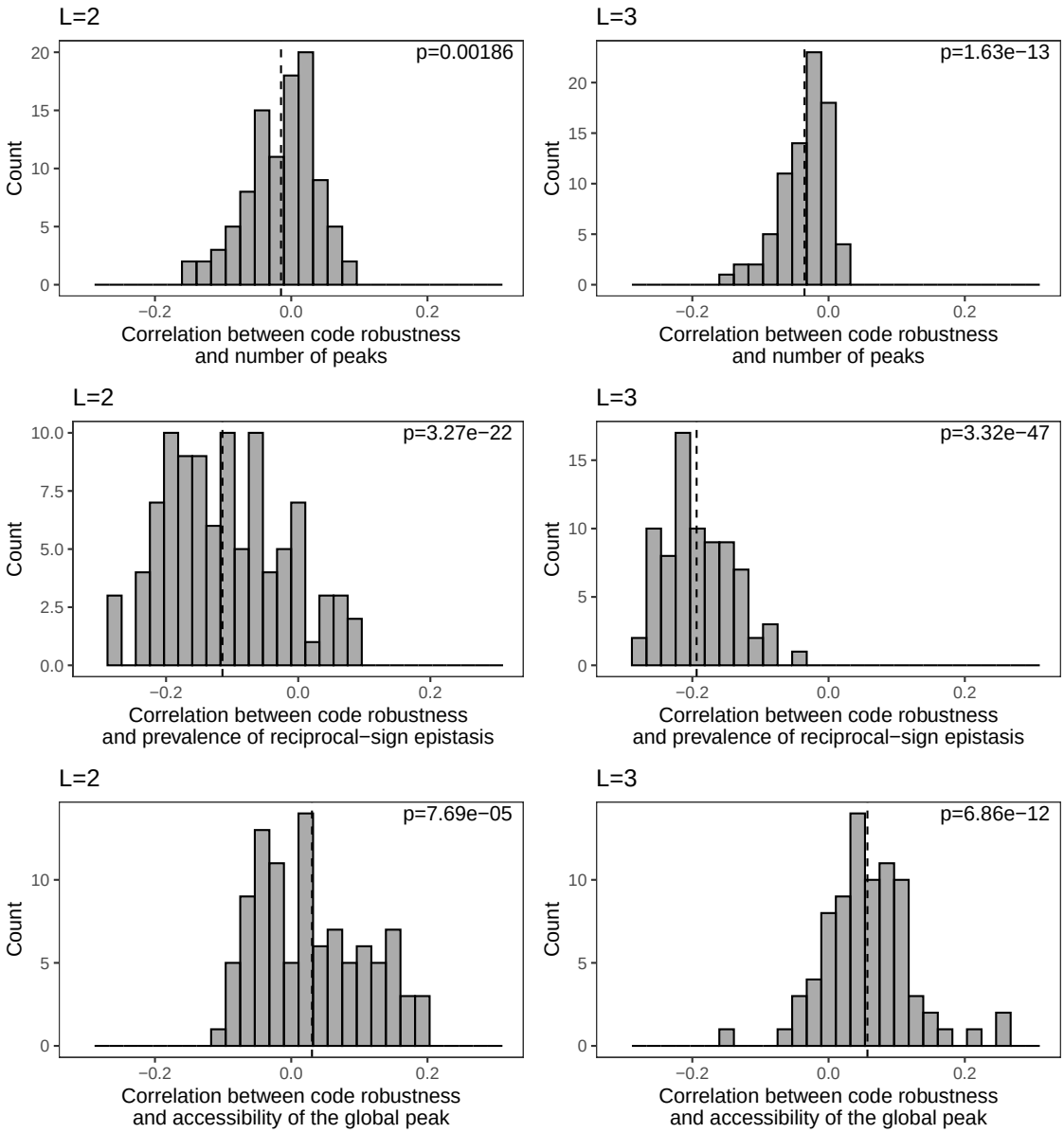


FIGURE S4.19: Distribution of Pearson's correlation coefficients between different measures of landscape ruggedness and code robustness, for 100 fitness landscapes with $L = 2$ (left column) and 80 fitness landscapes with $L = 3$ (right column). The rows, from top to bottom, correspond to the number of peaks in the landscape, the prevalence of reciprocal sign epistasis, and the accessibility of the global peak. In each histogram, the dashed line denotes the mean of the distribution. P-value in the upper right corner of each plot corresponds to a one-sample one-sided t-test of the mean being significantly different from 0 (significantly lower than 0 for number of peaks and proportion of reciprocal sign epistasis, significantly greater than 0 for accessibility of the global peak). Plots are centered on zero to emphasize the weight of each distribution.

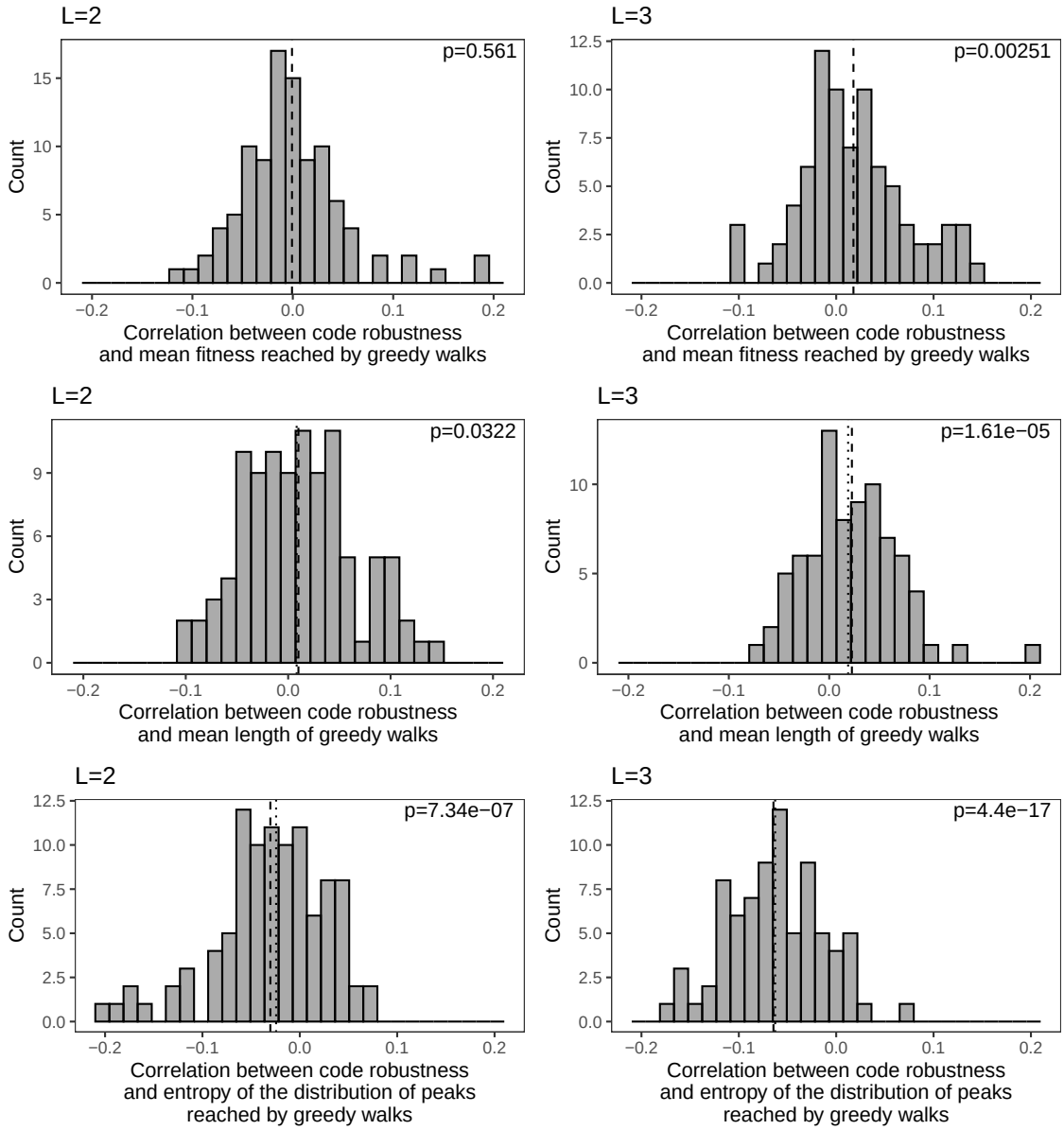


FIGURE S4.20: Distribution of Pearson's correlation coefficients between the outcomes of greedy adaptive walks and code robustness, for 100 fitness landscapes with $L = 2$ (left column) and 80 fitness landscapes with $L = 3$ (right column). The rows, from top to bottom, correspond to mean fitness, mean number of steps, and entropy of the distribution of reached peaks. In each histogram, the dashed line denotes the mean of the distribution. P-value in the upper right corner of each plot corresponds to a one-sample one-sided t-test of the mean being significantly different from 0 (significantly lower than 0 for entropy of the distribution of reached peaks, significantly greater than 0 for mean fitness and mean number of steps). Plots are centered on zero to emphasize the weight of each distribution.

S4.9 *Causes of the correlation between code robustness and mean fitness reached by greedy adaptive walks*

In our simulations of greedy adaptive walks, we observed a positive correlation between code robustness and mean fitness reached by the greedy adaptive walks in 4 out of 6 data sets (GB1, ParD-ParE2, ParD-ParE3, and ParB-*NBS*) and a negative one in the remaining two (ParB-*parS* and DHFR). What is the reason behind these conflicting results? To answer this question we investigated the subsets of landscapes that preserve the size of the global peak and in which the global peak consists of a single connected region in genotype space.

One possible explanation is that the landscapes differ in the relationship between code robustness and the mean height of fitness peaks. However, we observe that in all six data sets there is a negative correlation between code robustness and the mean height of fitness peaks ($R = -0.106$, $p = 7.20 \cdot 10^{-11}$, GB1; $R = -0.144$, $p < 2.2 \cdot 10^{-16}$, ParD-ParE2; $R = -0.0794$, $p = 5.81 \cdot 10^{-11}$, ParD-ParE3; $R = -0.168$, $p = 5.25 \cdot 10^{-16}$, ParB-*parS*; $R = -0.00318$, $p = 0.856$, ParB-*NBS*; $R = -0.075$, $p = 5.55 \cdot 10^{-10}$, DHFR).

The mean fitness reached by greedy adaptive walks depends not only on the heights of a landscape's fitness peaks, but also on the sizes of their basins of attraction, i.e., the number of greedy walks that terminate on them. Given the mostly positive relationship between code robustness and mean fitness reached by the greedy walks, we hypothesized that under robust codes the basins of attraction of the high-fitness peaks are relatively larger compared to those of less robust codes; in other words, that under robust codes the adaptive walks tend to converge on a smaller number of high-fitness peaks. Indeed, we observe that with increasing code robustness the Shannon entropy of the distribution of peaks reached by the greedy walks decreases in all six data sets (Supp. Tab. S4.4; Methods). To model the relationship between code robustness, peak height, and basin of attraction more explicitly, we fitted a linear model that, for each of the genetic codes and each data set, predicts the logarithm of the size of the basin of attraction of a peak as a linear function of its height:

$$\log(\text{size of basin}) = \beta_0 + \beta_1(\text{peak height}).$$

The β_1 coefficient controls how fast the size of the basin changes with peak height; for example, using the standard genetic code and the GB1 landscape, the coefficient is 0.591, meaning that if the peak height increases by 1, the size of the basin is expected to increase $\exp(0.591) \approx 1.8$-times. The bigger the β_1 coefficient, the faster the basin of attraction grows with peak height and the more concentrated the ends of the adaptive walks are on the high peaks. Having computed the β_1 coefficients for all genetic codes in the subset of genetic codes that preserve the size of the global peak, we then correlated them with the corresponding robustness. We observe the expected positive correlation in all data sets except for DHFR ($R = 0.141$, $p < 2.2 \cdot 10^{-16}$, GB1; $R = 0.250$, $p < 2.2 \cdot 10^{-16}$, ParD-ParE2; $R = 0.228$, $p < 2.2 \cdot 10^{-16}$, ParD-ParE3; $R = 0.0987$, $p = 2.13 \cdot 10^{-16}$, ParB-*parS*; $R = 0.117$, $p = 2.40 \cdot 10^{-11}$, ParB-*NBS*; but $R = -0.078$, $p = 1.10 \cdot 10^{-10}$, DHFR). The DHFR result is explained in more detail below. For the remaining five data sets, these analyses show that under robust genetic codes, evolutionary trajectories to adaptation become more predictable, in that they converge on a smaller number of adaptive peaks, and moreover, they preferentially converge on high-fitness peaks.

In summary, the relationship between code robustness and the mean fitness reached by greedy adaptive walks is influenced by two crucial factors: the mean height of peaks in the landscape and the sizes of their basins of attraction. While under robust genetic codes the peaks are on average lower, the high-fitness peaks have larger basins of attraction. This interplay between the average height of peaks and the sizes of their basins of attraction determines the sign of the correlation between code robustness and mean fitness reached by the greedy walks, with our analyses suggesting that in most cases, the larger basins of attraction of high-fitness peaks compensates for their reduced height.

S4.9.1 *Greedy walks in the DHFR landscape*

To understand the negative correlation between code robustness and mean fitness in the DHFR landscape, we further investigated the greedy adaptive walks under 100,000 amino acid permutation codes that preserve the number of codons per amino acid as in the standard genetic code (see also Supp. Section S4.6). We used this set of genetic codes to eliminate the potentially confounding factor of different number of codons per amino acid among the classical amino acid permutation codes. Also under these codes, we observe a negative correlation between code robustness and mean fitness (Supp. Tab. S4.18).

We observed that vast majority (99.3%) of the greedy walks terminates on local peaks with C, D, or E in their second position, with the XCX peaks having a significantly lower fitness than the XDX and XEX peaks (mean fitness -0.861 vs. 3.334, $p < 2.2 \cdot 10^{-16}$, Welch two sample t-test). Consequently, mean fitness reached by the greedy walks is strongly negatively correlated with the proportion of walks that terminate on an XCX peak ($R = -0.958$, $p < 2.2 \cdot 10^{-16}$). We thus hypothesized that the observed correlation between code robustness and mean fitness might be due to increased probability of reaching the XCX peaks under robust codes. Indeed, we see a positive correlation between code robustness and the proportion of greedy walks terminating at XCX peaks ($R = 0.060$, $p < 2.2 \cdot 10^{-16}$). To find out why, we divided the 100,000 amino acid permutation codes into two groups – those where the proportion of greedy walks terminating at XCX peaks is lower than 5% and those where the proportion is larger (Supp. Fig. S4.21) – and tried to identify whether these genetic codes differ systematically in any way. We observe that all the codes in the low XCX proportion group allow a C-to-D or C-to-E mutation (or both), while none of the codes in the high XCX proportion group do. The proportion of greedy walks that reach an XCX peak under a given genetic code is thus largely determined by whether the genetic code allows a C-to-D or C-to-E mutation. This is because under codes that allow a C-to-D or C-to-E mutation, the intermediate-fitness XCX sequences neighbor the high-fitness XDX/XEX sequences, and are thus less likely to be local peaks (0.572 vs. 4.98 local XCX peaks in the codes that do allow a C-to-D or C-to-E mutation and in codes that do not, resp., $p < 2.2 \cdot 10^{-16}$, Welch two sample t-test) and, if they are local peaks, their fitness is on average higher (0.963 vs. -0.305, $p < 2.2 \cdot 10^{-16}$, Welch two sample t-test) and their basin of attraction smaller (250.89 vs. 396.53 greedy walks, $p < 2.2 \cdot 10^{-16}$, Welch two sample t-test) than under codes that do not allow either of these two mutations. At the same time, we observe that robust codes are less likely to allow the C-to-D and C-to-E mutations (code robustness 0.342 vs. 0.345 for codes that do allow a C-to-D or C-to-E mutation and for codes that do not allow either of these mutations,

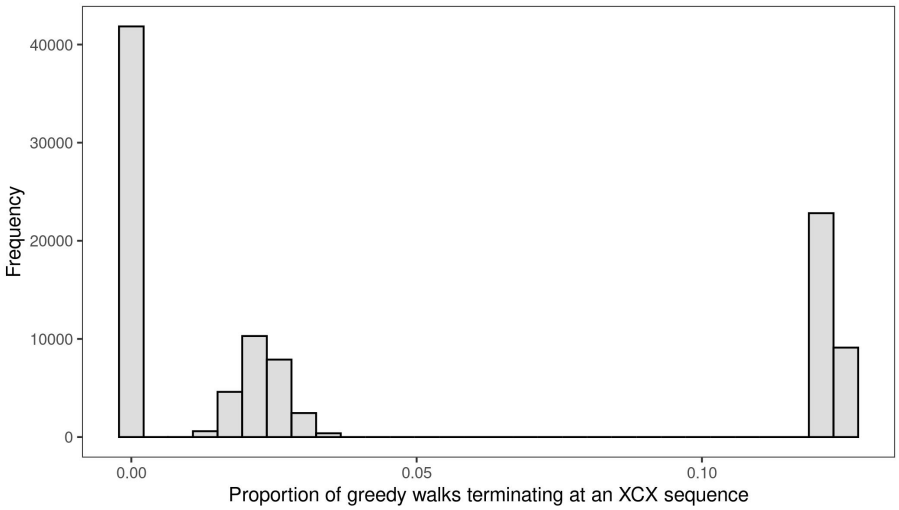


FIGURE S4.21: Histogram of the proportion of greedy walks terminating at an XCX sequence, in the set of 100,000 amino acid permutation codes restricted to preserve number of codons per amino acid as in the standard genetic code.

resp., $p < 2.2 \cdot 10^{-16}$, Welch two sample t-test), because both C-to-D and C-to-E mutations are considered non-conservative (Supp. Fig. S4.3).

To summarize, the observed negative correlation between code robustness and mean fitness observed in the DHFR landscape can be explained by a particular idiosyncracy in the landscape, in which the mean fitness is largely determined by whether the genetic code allows a C-to-D or C-to-E mutation. The negative correlation arises because these mutations are less likely to be allowed under robust codes.

S4.10 *Weak mutation adaptive walks*

In the main text, we modeled evolution using greedy adaptive walks, which correspond to a strong-mutation-strong-selection regime. To understand how our results depend on population genetic conditions, we also performed weak mutation adaptive walks. The weak mutation adaptive walks represent adaptive evolution under the regime where mutations occur so infrequently that any mutation will either go to extinction or to fixation prior to the arrival of a subsequent mutation. The probability of fixation depends on both the improvement in fitness and the population size, which controls the strength of genetic drift. For each genetic code, each landscape and each choice of one of four different population sizes, we simulated 100,000 random walks, initialized in randomly chosen mRNA sequences. In each subsequent step, a random single-nucleotide mutation was proposed, and accepted with probability

$$P_{\text{accept}} = \begin{cases} \frac{1-\exp(f_{\text{old}}-f_{\text{new}})}{1-\exp(N(f_{\text{old}}-f_{\text{new}}))} & \text{if } f_{\text{old}} \neq f_{\text{new}} \\ \frac{1}{N} & \text{otherwise} \end{cases}$$

where f_{old} denotes the fitness of the current genotype, f_{new} the fitness of the proposed genotype, and N is the population size. This corresponds to the exact fixation probability under the Moran process [292]. We ran each adaptive walk for 500 steps, using population sizes $N \in \{10; 100; 10,000; 1,000,000\}$.

We obtained qualitatively the same results as for the greedy adaptive walks for the set of 100,000 amino acid permutation codes (Supp. Tab. S4.21), as well as for the Ostrov codes (Supp. Tab. S4.22, S4.23, and S4.24).

	GB1	ParD-ParE2	ParD-ParE3	ParB-*parS*	ParB-*NBS*	DHFR
$N = 10$	0.120	0.124	0.050	-0.058, $p = 0.0057$	0.0891	-0.123
$N = 100$	0.096	0.155	0.029, $p = 0.018$	-0.081	0.121	-0.123
$N = 10,000$	0.091	0.148	0.018, $p = 0.128$	-0.083	0.119	-0.129
$N = 1,000,000$	0.091	0.148	0.018, $p = 0.131$	-0.083	0.119	-0.128

TABLE S4.21: Correlation of code robustness with the mean fitness reached after 500 steps of the weak mutation adaptive walks, for different population sizes N, using the amino acid permutation codes. Correlations are statistically significant unless specified otherwise. Data pertain to the subset of amino acid permutation codes that preserve the size of the global peak and under which the global peak forms a single connected region in the genotype space.

	GB1	ParD-ParE2	ParD-ParE3	ParB-*parS*	ParB-*NBS*	DHFR
$N = 10$	0.300	0.295	-0.141	0.170	0.111	-0.186
$N = 100$	0.268	0.380	-0.050	0.148	-0.207	-0.194
$N = 10,000$	0.306	0.410	0.009, $p = 0.0056$	0.159	-0.238	-0.174
$N = 1,000,000$	0.307	0.410	0.010, $p = 0.0014$	0.160	-0.239	-0.173

TABLE S4.22: Correlation of code robustness with the mean fitness reached after 500 steps of weak mutation adaptive walks, for different population sizes N, using the Ostrov codes. All correlations are statistically significant unless stated otherwise. All results pertain to the subset of codes that preserve the size of the global peak and under which the global peak forms a single connected region in the genotype space.

	GB1	ParD-ParE2	ParD-ParE3	ParB-*parS*	ParB-*NBS*	DHFR
$N = 10$	-0.117	-0.200	-0.087	-0.064	-0.262	0.221
$N = 100$	-0.014, $p = 0.294$	-0.160	0.049	0.0025, $p = 0.763$	-0.056	0.224
$N = 10,000$	0.104	-0.006, $p = 0.428$	0.185	0.058	0.013, $p = 0.170$	0.227
$N = 1,000,000$	0.105	-0.003, $p = 0.703$	0.187	0.058	0.014, $p = 0.144$	0.227

TABLE S4.23: Correlation of the number of split codon blocks with the mean fitness reached after 500 steps of weak mutation adaptive walks, for different population sizes N, using the Ostrov codes. All correlations are statistically significant unless specified otherwise. All results pertain to the subset of codes that preserve the size of the global peak and under which the global peak forms a single connected region in the genotype space.

	Number of split codon blocks	GB1	ParD-ParE2	ParD-ParE3	ParB-*parS*	ParB-*NBS*	DHFR
$N = 10$	0	-0.255	-0.088, $p = 0.00268$	-0.044, $p = 0.037$	-0.031, $p = 0.450$	-0.107, $p = 0.011$	-0.677
	1	-0.145	-0.072	-0.008, $p = 0.337$	-0.013, $p = 0.464$	-0.157	-0.291
	2	-0.084	-0.050	0.011, $p = 0.040$	0.004, $p = 0.752$	-0.159	-0.184
	3	-	-	0.016, $p = 0.0017$	-	-	-0.110
$N = 100$	0	-0.376	-0.229	-0.077, $p = 2.42 \cdot 10^{-4}$	-0.029, $p = 0.480$	-0.048, $p = 0.253$	-0.575
	1	-0.246	-0.145	-0.033	0.0056, $p = 0.745$	-0.079	-0.246
	2	-0.149	-0.077	0.0011, $p = 0.828$	0.025, $p = 0.043$	-0.082	-0.163
	3	-	-	0.021	-	-	-0.102
$N = 10,000$	0	-0.769	-0.765	-0.565	-0.080, $p = 0.051$	-0.108, $p = 0.010$	-0.760
	1	-0.631	-0.626	-0.495	-0.023, $p = 0.179$	-0.104	-0.385
	2	-0.458	-0.432	-0.402	0.013, $p = 0.281$	-0.086	-0.263
	3	-	-	-0.275	-	-	-0.170
$N = 1,000,000$	0	-0.773	-0.634	-0.573	-0.081, $p = 0.048$	-0.109	-0.749
	1	-0.635	-0.771	-0.504	-0.024, $p = 0.169$	-0.104	-0.381
	2	-0.463	-0.440	-0.411	0.013, $p = 0.290$	-0.086	-0.261
	3	-	-	-0.282	-	-	-0.169

TABLE S4.24: Correlation of the number of stop codons, conditioned on the number of split codon blocks, with the mean fitness reached after 500 steps of weak mutation adaptive walks, using the Ostrov codes. Results for 4 split codon blocks not shown because all codes with 4 split codon blocks have 2 stop codons. All correlations are statistically significant, unless specified otherwise. All results pertain to the subset of codes that preserve the size of the global peak and under which the global peak forms a single connected region in the genotype space.

S4.11 *Epistasis under the Ostrov codes*

The Ostrov codes can vary in the number of split codon blocks and the number of stop codons they contain. The effect of increasing the number of split codon blocks on the prevalence of the different forms of epistasis follows intuition: As the number of split codon blocks increases, the number of synonymous mutations decreases, thus decreasing the prevalence of squares with no epistasis. And among the epistatic squares, increasing the number of split codon blocks increases the prevalence of simple and reciprocal sign epistasis (Supp. Tab. S4.8). However, the effect of increasing the number of stop codons on the different forms of epistasis is more complicated (Supp. Tab. S4.10). First, we observe a strong positive correlation between the number of stop codons and the prevalence of no epistasis. This is because the more stop codons a genetic code has, the more mRNAs contain at least one stop codon, and hence the more squares that consist entirely of sequences containing stop codons. As we have assigned the same fitness value to all sequences containing stop codons (Methods), these squares will be classified as exhibiting no epistasis. Second, contrary to expectation, we observe a strong positive correlation between the number of stop codons and the prevalence of magnitude epistasis, and a strong negative correlation between the number of stop codons and simple and reciprocal sign epistasis (Supp. Tab. S4.10).

To understand these results, we must think of the types of squares that mRNA sequences containing stop codons can be part of. Not considering the trivial squares consisting only of sequences containing at least one stop codon, there are 6 possible configurations, which we depict in Fig. S4.22. In the following, we will mostly focus on the configurations in Panels A and B, and discuss the remaining 4 configurations at the end. Configuration A involves squares where the "wild type" sequence contains one stop codon and one of the mutations changes the stop codon to a sense codon, while the second mutation happens in any of the remaining codons (Fig. S4.22A). Configuration B involves squares where the wild type sequence contains one stop codon and both mutations change the stop codon to a sense codon (Fig. S4.22B). In both cases we assume that the double mutant does not contain any stop codons. To understand the influence these squares have on the prevalence of epistasis, we need to know what types of epistasis squares A and B can exhibit and how many of these squares there are.

All the type A squares show magnitude epistasis, regardless of the fitness values of the two sequences without stop codons. How many squares of this type are there? While the exact number will depend on the particular location of the stop codons in the genetic code, we can estimate the number to be roughly

$$N_A = L \cdot n_{\text{STOP}} \cdot n_{\text{sense}}^{L-1} \cdot n_{\text{STOP}\to\text{sense}} \cdot (L-1) n_{\text{sense}\to\text{sense}},$$

where L is the total number of codons in the sequence, n_{STOP} is the number of stop codons in the code, n_{sense} is the number of sense codons in the code, $n_{\text{STOP}\to\text{sense}}$ is the expected number of mutations from a stop codon to a sense codon, and $n_{\text{sense}\to\text{sense}}$ is the expected number of mutations from a sense codon to another sense codon. The first three terms quantify the number of sequences containing exactly one stop codon, while the fourth and fifth terms quantify the number of possible mutations that would give rise to a type A square.

For the type B squares, the type of epistasis depends on the exact fitness values of the three sequences that do not contain stop codons, which can cause the square to exhibit magnitude, simple sign, or reciprocal sign epistasis. Using reasoning similar to that above, we estimate the number of these squares as roughly

$$N_B = L \cdot n_{\text{STOP}} \cdot n_{\text{sense}}^{L-1} \cdot n_{\text{STOP}\to\text{sense}} \cdot n_{\text{STOP}\to\text{sense}}.$$

Even without taking into account the fact that the two mutations must happen in two different nucleotide positions, and hence the last two terms are at most $9 \cdot 6$, N_A is clearly at least $(L-1)$-times bigger than N_B. In other words, it is much more likely that the second mutation happens in a different codon, than that both mutations happen in the stop codon. As all type A squares exhibit magnitude epistasis, we would thus expect that the prevalence of magnitude epistasis, relative to simple sign and reciprocal sign epistasis, increases as the number of stop codons increases.

Do the remaining possible configurations (Panels C and D) change the result? All squares in Panel C exhibit magnitude epistasis and will thus further increase the prevalence of magnitude epistasis. On the other hand, squares in Panel D, consisting of a wild type and a double mutant that do contain a stop codon and two single mutants that do not, always exhibit reciprocal sign epistasis. However, in the Ostrov codes the number of such squares is extremely low, due to the fact that the stop codons can be placed in only a small handful of positions; in fact, the maximum number of type D squares in the whole landscape is 2 (Fig. S4.22D), so their effect on reciprocal sign epistasis is negligible.

To conclude, contrary to expectation, increasing the number of stop codons increases the prevalence of magnitude epistasis, relative to simple sign and reciprocal sign epistasis, because the number of squares with two neighboring sequences containing stop codons (Fig. S4.22A and C) is much larger than the number of squares where a low-fitness sequence containing a stop codon separates two higher-fitness variants without stop codons (Fig. S4.22B and D).

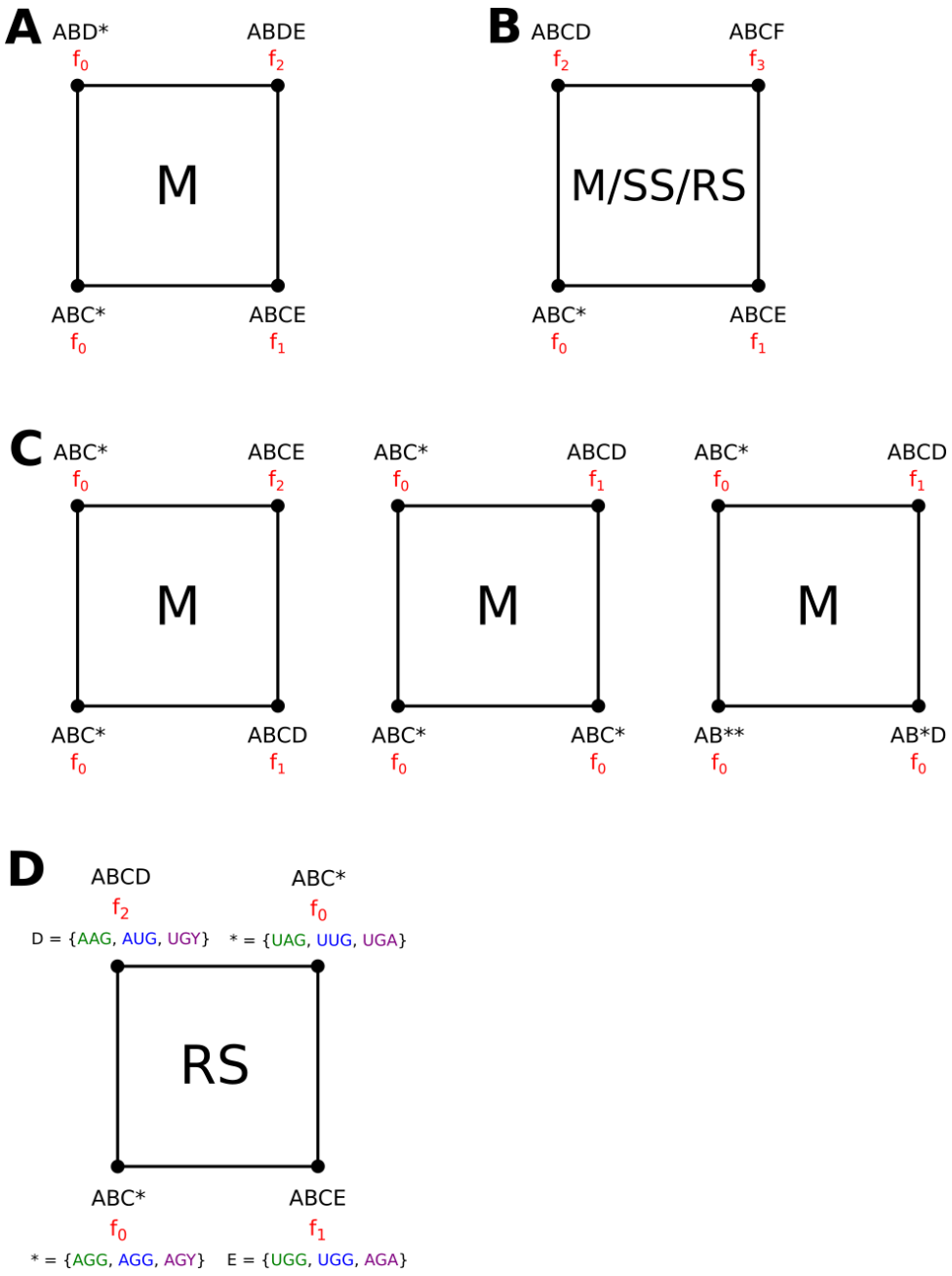


FIGURE S4.22: Possible configurations of squares involving at least 1 and at most 3 sequences containing stop codons, for sequences of length $L = 4$, with the stop codon occupying the last position. A, B, C, D, E, and F denote arbitrary amino acids, not necessarily different from each other. Fitness values are denoted by $f_0, \ldots, f_3$, with $f_0 < f_i$ for $i = 1, 2, 3$, while we assume no particular ordering of f_1, f_2, and f_3. The letters in the middle of the squares denote the possible types of epistasis a given configuration can exhibit; M = magnitude epistasis, SS = simple sign epistasis, RS = reciprocal sign epistasis. In D, possible assignments of the last codon, based on the Ostrov codes, are listed in green, blue, and violet; notice that the green and blue assignments are not compatible with the violet one, as AGG and AGA codons need to be assigned the same amino acid (or stop signal), and cannot thus at the same time encode a stop signal (green, blue) and an amino acid (violet). Thus, the maximum number of squares of type D in a landscape caused by an Ostrov code is 2.

S4.12 *Examples of Ostrov codes promoting or diminishing evolvability*

In the main text we discussed common features of Ostrov codes enhancing or diminishing evolvability. Namely, we observed that genetic codes promoting evolvability tend to have fewer split codon blocks, fewer stop codons, and higher physicochemical robustness, while the opposite is true for genetic codes that diminish evolvability, with the exception of the number of stop codons (Fig. 4.6 and Supp. Fig. S4.14). Here, we present concrete examples of genetic codes following these design principles.

Panels A and B in Supp. Fig. S4.23 show the two genetic codes with the highest robustness of the consistently high-ranking Ostrov codes ("Evolvable Ostrov Code A" and "Evolvable Ostrov Code B"). They have a robustness of 0.385, the minimal number of zero split codon blocks, and the minimal number of two stop codons. Engineering these codes in a living organism is in principle possible with existing technology, although it requires the reassignment of five of the seven freed codons, which is no small feat. In contrast, most experimental studies of rewired genetic codes only change the meaning of the UAG stop codon [222]. Bacterial strains containing no genomic TAG, as well as a variety of orthologous translation systems that decode UAG as a nonstandard amino acid are commercially available, so engineering a strain with reassigned UAG is relatively straightforward. The best of such codes, in terms of mean fitness reached in our simulations, is depicted in Supp. Fig. S4.23C ("Evolvable Ostrov Code C"). It reassigns UAG to glutamate, and it ranks among the best 25% of Ostrov codes for all data sets except for GB1 (Supp. Tab. S4.13).

How do these codes compare to the standard genetic code in our evolutionary simulations? Evolvable Ostrov Code A ranks better in terms of fitness on five out of six data sets, and Evolvable Ostrov Codes B and C rank better than the standard genetic code in four out of six cases (Supp. Tab. S4.13). It is also worth noting that in four out of six data sets, the mean fitness reached under the standard genetic code, as well as Evolvable Ostrov Codes A, B, and C is higher than when using genetic codes specifically designed for increased evolvability [93] (Supp. Tab. S4.13), even though they are much easier to engineer than those proposed by Pines et al. [93]. Interestingly, for the ParB-*NBS* data set, some of the codes proposed by Pines et al. perform worse than any of the 194,481 Ostrov codes. This further confirms that decreasing code robustness does not in general lead to an increase in evolvability.

While for the genetic codes promoting evolvability it is possible to optimize all three design principles at once, this is not possible for the codes that diminish evolvability. For example, a genetic code where all free codon blocks are assigned to a stop signal will have the maximum possible number of stop codons (9), but it will also have the minimal number of split codon blocks (0). It is thus impossible to highlight one genetic code that would be expected to decrease evolvability the most based on our design principles. Instead, in Panel D of Fig. S4.23 we show a genetic code that ranks among the bottom 12.5% of codes for all six data sets (Supp. Tab. S4.13). We again compared the level to which this code diminishes evolvability with codes specifically designed to slow down the rate of evolution [195] (Supp. Tab. S4.13). While reducing evolvability beyond the majority of the Ostrov codes, the mean fitness reached in adaptive walks using genetic code D is still much higher than when using the codes proposed by Calles et al. [195]. However, we emphasize that, similar to the codes proposed by Pines et al. [93], the codes proposed by Calles et al. require extensive genome recoding, such that the majority of codons are "null", meaning

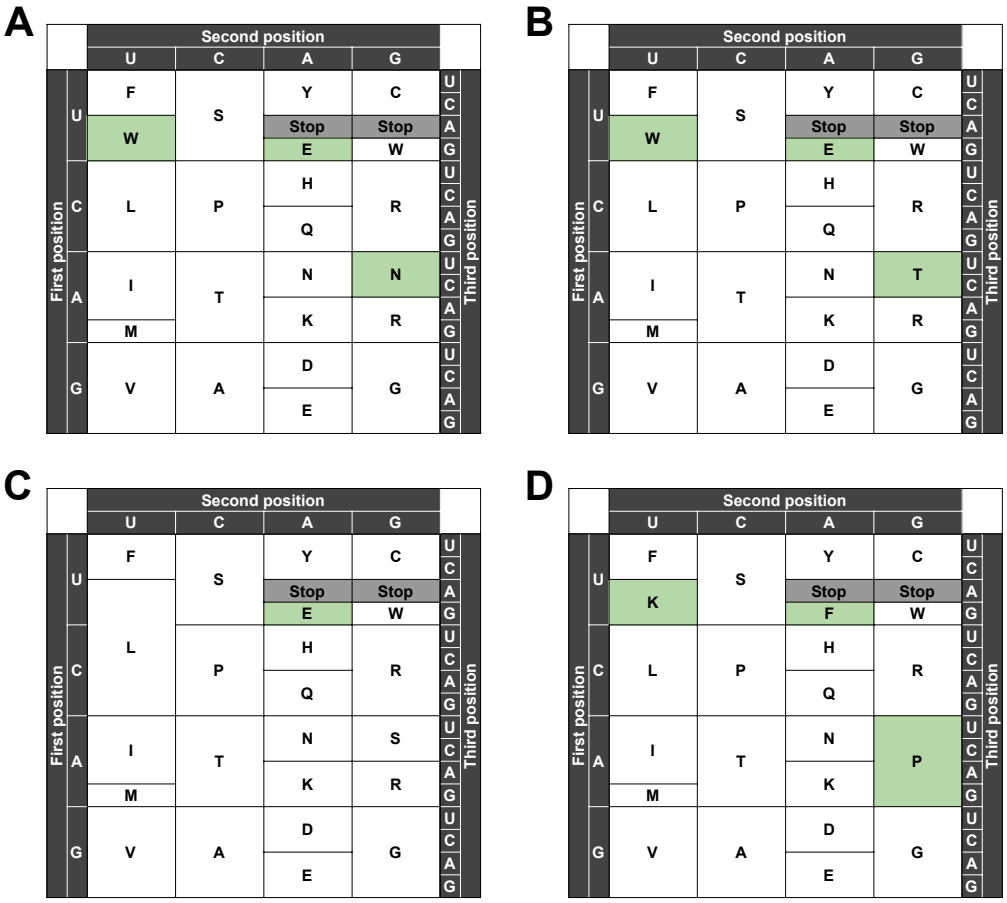


FIGURE S4.23: Examples of codes promoting (A-C) or diminishing (D) evolvability, identified based on their robustness, number of stop codons, and number of split codon blocks, as well as the results of the greedy adaptive walks. Changes compared to the standard genetic code are highlighted in green.

they encode neither an amino acid nor a stop signal. We hope the design principles we have identified here will provide guidance for engineering genetic codes that significantly enhance or diminish evolvability, but remain within reach of current technology.

5

CONCLUSION

Language is the means by which humans translate their thoughts into meaningful messages for others. Similarly, the genetic code is the means by which cells translate information encoded in nucleic acids into functional proteins. The genetic code shares many properties with human languages: It contains the blueprint of its own history as well as the history of life, akin to how languages are shaped by the histories of their nations. Using different genetic codes, like speaking different languages, impedes communication between cells. On the other hand, sharing a genetic code facilitates horizontal gene transfer but also makes cells susceptible to genetic parasites, similarly to how people speaking the same language can share not only useful information but also misleading ideas or propaganda. And just as certain concepts may be more easily expressed in one language than another, some evolutionary paths are more likely to occur under one genetic code than another. The main "dialect" of this molecular language, the standard genetic code, was deciphered 60 years ago. Yet, to this day, we continue to explore its origins, its ongoing evolution, and its impact on evolutionary processes. In this thesis, we have contributed to these efforts in three distinct areas.

In Chapter 2, we focused on the history of the genetic code and the insights we can gain from its structure. We demonstrated the importance of caution when interpreting the seemingly non-random patterns in the genetic code, as numerous confounding factors exist, and the way in which the null set of rewired genetic codes is constructed can significantly influence the results. Ideally, we would construct the null set to accurately represent the potential genetic codes that life might have converged on instead of the standard genetic code. However, this would require knowledge of the exact conditions and limitations that influenced the evolution of the standard genetic code. For example, is the exact synonymous block structure of the standard code crucial or could it look differently? Is it important how many codons encode a given amino acid? Is there any reason for the three stop codons being specifically UAG, UAA, and UGA? And why are there even three stop codons if, in theory, the job could be as well done by just one? We do not know. One possible way forward is to study extant alternative genetic codes, which may provide insights into the constraints on the genetic code and the way it evolves. However, the extent of what we can learn from this source of information is limited, as the current evolution of the genetic code occurs under conditions that are vastly different from those at its origins.

Even though their informativeness regarding the origins of the standard genetic code is limited, extant alternative genetic codes are far from mere curiosities. By studying exceptions to the rules, we gain a deeper understanding of the rules themselves. The alternative genetic codes can thus enhance our understanding of the basic molecular process of translation. Moreover, investigating alternative genetic codes helps us understand the mechanisms behind their evolution, their prevalence, and their possible functions. To achieve this, it is crucial to identify how many and which genetic codes exist, ideally through large-scale computational screens of available genomes. In Chapter 3, we conducted such a screen on more than 15.5 million viral genomes, uncovering eight new instances of genetic code alteration. Notably, four of these alterations only pertain

to part of the corresponding genome, while the rest of the genome uses the standard genetic code. The switch between the two genetic codes triggers the transition to the late phase of viral infection, and the genetic code thus becomes a mechanism for gene regulation.

The genetic code can also influence protein functionality and organism fitness indirectly, through its impact on evolutionary processes. Over the past 15 years, we have begun to understand that the outcomes of adaptive evolution are shaped by mutation bias, the relative differences in mutation rates between different types of mutations [178], [293], [294]. The genetic code imposes a form of mutation bias on amino acid pairs, where some pairs can be connected by a single-nucleotide substitution, while others require two or even three. In Chapter 4, we show that evolutionary outcomes can vary significantly within the same fitness landscape under different genetic codes. Specifically, we demonstrate that robust genetic codes, which predominantly allow substitutions between physicochemically similar amino acids, promote evolvability. However, many questions concerning the relationship between the genetic code and evolution remain. Does the structure of the genetic code manifest itself in the spectrum of adaptive substitutions, leading to more frequent substitutions between better-connected amino acid pairs, similar to how more prevalent types of single-nucleotide mutations appear more frequently among adaptive changes [178], [294]? Are amino acids encoded by larger codon blocks more abundant in proteomes of living organisms simply because they are easier to find [232]? And can the structure of the genetic code teach us anything about codon bias [133]? The influence of the genetic code on evolutionary processes has been underappreciated so far, perhaps because almost all protein evolution on Earth occurs under the standard genetic code, making its influence ubiquitous. Future studies, both computational and experimental, should aim to address this knowledge gap.

Just as language shapes and is shaped by human culture and history, the genetic code shapes and was shaped by the evolution of life. That is why, even though it is merely a simple mapping of 64 codons to amino acids, it can still, after decades of research, teach us about the history of life, the principles of basic cellular processes, and the intricate mechanisms of evolution.

BIBLIOGRAPHY

[1] R. D. Knight, S. J. Freeland, and L. F. Landweber, "Rewiring the keyboard: Evolvability of the genetic code", *Nature Reviews Genetics*, vol. 2, no. 1, pp. 49–58, 2001.

[2] C. R. Woese, D. H. Dugre, W. C. Saxinger, and S. A. Dugre, "The molecular basis for the genetic code.", *Proceedings of the National Academy of Sciences*, vol. 55, no. 4, pp. 966–974, 1966.

[3] E. V. Koonin, "Comparative genomics, minimal gene-sets and the last universal common ancestor", *Nature Reviews Microbiology*, vol. 1, no. 2, pp. 127–136, 2003.

[4] M. Ibba, H. D. Becker, C. Stathopoulos, D. L. Tumbula, and D. Söll, "The adaptor hypothesis revisited", *Trends in Biochemical Sciences*, vol. 25, no. 7, pp. 311–316, 2000.

[5] T. Muramatsu, K. Nishikawa, F. Nemoto, *et al.*, "Codon and amino-acid specificities of a transfer RNA are both converted by a single post-transcriptional modification", *Nature*, vol. 336, no. 6195, pp. 179–181, 1988.

[6] V. Perret, A. Garcia, H. Grosjean, J.-P. Ebel, C. Florentz, and R. Giegé, "Relaxation of a transfer RNA specificity by removal of modified nucleotides", *Nature*, vol. 344, no. 6268, pp. 787–789, 1990.

[7] C. R. Woese, "On the evolution of the genetic code", *Proceedings of the National Academy of Sciences*, vol. 54, no. 6, pp. 1546–1552, 1965.

[8] F. Crick, "The origin of the genetic code", *Journal of Molecular Biology*, vol. 38, no. 3, pp. 367–379, 1968.

[9] D. Haig and L. D. Hurst, "A quantitative measure of error minimization in the genetic code", *Journal of Molecular Evolution*, vol. 33, pp. 412–417, 1991.

[10] S. J. Freeland and L. D. Hurst, "The genetic code is one in a million", *Journal of Molecular Evolution*, vol. 47, pp. 238–248, 1998.

[11] R. Geyer and A. Madany Mamlouk, "On the efficiency of the genetic code after frameshift mutations", *PeerJ*, vol. 6, e4825, 2018.

[12] L. Bartonek, D. Braun, and B. Zagrovic, "Frameshifting preserves key physicochemical properties of proteins", *Proceedings of the National Academy of Sciences*, vol. 117, no. 11, pp. 5907–5912, 2020.

[13] E. Firnberg and M. Ostermeier, "The genetic code constrains yet facilitates Darwinian evolution", *Nucleic Acids Research*, vol. 41, no. 15, pp. 7420–7428, 2013.

[14] E. V. Koonin and A. S. Novozhilov, "Origin and evolution of the universal genetic code", *Annual Review of Genetics*, vol. 51, no. 1, pp. 45–62, 2017.

[15] K. Vetsigian, C. Woese, and N. Goldenfeld, "Collective evolution and the genetic code", *Proceedings of the National Academy of Sciences*, vol. 103, no. 28, pp. 10 696–10 701, 2006.

[16] C. R. Woese, "The fundamental nature of the genetic code: Prebiotic interactions between polynucleotides and polyamino acids or their derivatives", *Proceedings of the National Academy of Sciences*, vol. 59, no. 1, pp. 110–117, 1968.

[17] M. Yarus, "Amino acids as RNA ligands: A direct-RNA-template theory for the code's origin", *Journal of Molecular Evolution*, vol. 47, no. 1, pp. 109–117, 1998.

[18] M. Yarus, "RNA-ligand chemistry: A testable source for the genetic code", *RNA*, vol. 6, pp. 475–484, 2000.

[19] M. Yarus, J. G. Caporaso, and R. Knight, "Origins of the genetic code: The escaped triplet theory", *Annual Review of Biochemistry*, vol. 74, pp. 179–198, 2005.

[20] M. Yarus, J. J. Widmann, and R. Knight, "RNA–amino acid binding: A stereochemical era for the genetic code", *Journal of Molecular Evolution*, vol. 69, no. 5, pp. 406–429, 2009.

[21] J. T.-F. Wong, "A co-evolution theory of the genetic code", *Proceedings of the National Academy of Sciences*, vol. 72, no. 5, pp. 1909–1912, 1975.

[22] M. Di Giulio and M. Medugno, "The historical factor: The biosynthetic relationships between amino acids and their physicochemical properties in the origin of the genetic code", *Journal of Molecular Evolution*, vol. 46, no. 6, pp. 615–621, 1998.

[23] M. Di Giulio and M. Medugno, "The robust statistical bases of the coevolution theory of genetic code origin", *Journal of Molecular Evolution*, vol. 50, no. 3, pp. 258–263, 2000.

[24] M. Di Giulio, "Genetic code origin: Are the pathways of type Glu-tRNA$^{\mathrm{Gln}} \rightarrow$ Gln-tRNA$^{\mathrm{Gln}}$ molecular fossils or not?", *Journal of Molecular Evolution*, vol. 55, no. 5, pp. 616–622, 2002.

[25] R. Amirnovin, "An analysis of the metabolic theory of the origin of the genetic code", *Journal of Molecular Evolution*, vol. 44, no. 5, pp. 473–476, 1998.

[26] T. A. Ronneberg, L. F. Landweber, and S. J. Freeland, "Testing a biosynthetic theory of the genetic code: Fact or artifact?", *Proceedings of the National Academy of Sciences*, vol. 97, no. 25, pp. 13 690–13 695, 2000.

[27] D. H. Ardell, "On error minimization in a sequential origin of the standard genetic code", *Journal of Molecular Evolution*, vol. 47, 1–13, 1998.

[28] D. Haig and L. D. Hurst, "A quantitative measure of error minimization in the genetic code", *Journal of Molecular Evolution*, vol. 49, p. 708, 1999.

[29] S. J. Freeland, R. D. Knight, L. F. Landweber, and L. D. Hurst, "Early Fixation of an Optimal Genetic Code", *Molecular Biology and Evolution*, vol. 17, no. 4, pp. 511–518, 2000.

[30] D. Gilis, S. Massar, N. J. Cerf, and M. Rooman, "Optimality of the genetic code with respect to protein stability and amino-acid frequencies", *Genome Biology*, vol. 2, no. 11, research0049.1, 2001.

[31] H. Goodarzi, H. Shateri Najafabadi, H. A. Nejad, and N. Torabi, "The impact of including tRNA content on the optimality of the genetic code", *Bulletin of Mathematical Biology*, vol. 67, no. 6, 1355–1368, 2005.

[32] A. S. Novozhilov, Y. I. Wolf, and E. V. Koonin, "Evolution of the genetic code: Partial optimization of a random code for robustness to translation error in a rugged fitness landscape", *Biology Direct*, vol. 2, no. 1, p. 24, 2007.

[33] T. Butler, N. Goldenfeld, D. Mathew, and Z. Luthey-Schulten, "Extreme genetic code optimality from a molecular dynamics calculation of amino acid polar requirement", *Physical Review E*, vol. 79, no. 6, p. 060 901, 2009.

[34] A. S. Novozhilov and E. V. Koonin, "Exceptional error minimization in putative primordial genetic codes", *Biology Direct*, vol. 4, no. 1, p. 44, 2009.

[35] S. E. Massey, "A neutral origin for error minimization in the genetic code", *Journal of Molecular Evolution*, vol. 67, no. 5, p. 510, 2008.

[36] Y. I. Wolf and E. V. Koonin, "On the origin of the translation system and the genetic code in the RNA world by means of natural selection, exaptation, and subfunctionalization", *Biology Direct*, vol. 2, no. 1, p. 14, 2007.

[37] P. G. Higgs, "A four-column theory for the origin of the genetic code: Tracing the evolutionary pathways that gave rise to an optimized code", *Biology Direct*, vol. 4, no. 1, p. 16, 2009.

[38] E. Lerat, V. Daubin, H. Ochman, and N. A. Moran, "Evolutionary origins of genomic repertoires in bacteria", *PLOS Biology*, vol. 3, no. 5, e130, 2005.

[39] T. J. Treangen and E. P. C. Rocha, "Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes", *PLOS Genetics*, vol. 7, no. 1, pp. 1–12, 2011.

[40] J. F. Zürcher, W. E. Robertson, T. Kappes, *et al.*, "Refactored genetic codes enable bidirectional genetic isolation", *Science*, vol. 378, no. 6619, pp. 516–523, 2022.

[41] A. Nyerges, S. Vinke, R. Flynn, *et al.*, "A swapped genetic code prevents viral infections and gene transfer", *Nature*, vol. 615, no. 7953, pp. 720–727, 2023.

[42] N. Aggarwal, A. V. Bandhu, and S. Sengupta, "Finite population analysis of the effect of horizontal gene transfer on the origin of an universal and optimal genetic code", *Physical Biology*, vol. 13, no. 3, p. 036 007, 2016.

[43] A. Ambrogelly, S. Palioura, and D. Söll, "Natural expansion of the genetic code", *Nature Chemical Biology*, vol. 3, pp. 29–35, 2007.

[44] Y. Shulgina, "The evolution of alternative genetic codes", Ph.D. dissertation, Harvard University, 2022.

[45] M. Li and A. Tzagoloff, "Assembly of the mitochondrial membrane system: sequences of yeast mitochondrial valine and an unusual threonine tRNA gene", *Cell*, vol. 18, no. 1, pp. 47–53, 1979.

[46] G. Clark-Walker and G. Weiller, "The structure of the small mitochondrial DNA of *Kluyveromyces thermotolerans* is likely to reflect the ancestral gene order in fungi", *Journal of Molecular Evolution*, vol. 38, no. 6, pp. 593–601, 1994.

[47] D. Su, A. Lieberman, B. F. Lang, M. Simonović, D. Söll, and J. Ling, "An unusual $tRNA^{Thr}$ derived from $tRNA^{His}$ reassigns in yeast mitochondria the CUN codons to threonine", *Nucleic Acids Research*, vol. 39, no. 11, pp. 4866–4874, 2011.

[48] V. N. Gladyshev and G. V. Kryukov, "Evolution of selenocysteine-containing proteins: Significance of identification and functional characterization of selenoproteins", *BioFactors*, vol. 14, no. 1-4, pp. 87–92, 2001.

[49] D. L. Hatfield and V. N. Gladyshev, "How selenium has altered our understanding of the genetic code", *Molecular and Cellular Biology*, vol. 22, no. 11, pp. 3565–3576, 2002.

[50] G. Srinivasan, C. M. James, and J. A. Krzycki, "Pyrrolysine encoded by UAG in Archaea: Charging of a UAG-decoding specialized tRNA", *Science*, vol. 296, no. 5572, pp. 1459–1462, 2002.

[51] J.-F. Brugère, J. F. Atkins, P. W. O'Toole, and G. Borrel, "Pyrrolysine in Archaea: a 22nd amino acid encoded through a genetic code expansion", *Emerging Topics in Life Sciences*, vol. 2, no. 4, pp. 607–618, 2018.

[52] J. M. Tharp, A. Ehnbom, and W. R. Liu, "$tRNA^{Pyl}$: Structure, function, and applications", *RNA Biology*, vol. 15, no. 4-5, pp. 441–452, 2018.

[53] J. W. Roberts and J. Carbon, "Molecular mechanism for missense suppression in *E. coli*", *Nature*, vol. 250, no. 5465, pp. 412–414, 1974.

[54] C Citti, L Maréchal-Drouard, C Saillard, J. H. Weil, and J. M. Bové, "*Spiroplasma citri* UGG and UGA tryptophan codons: Sequence of the two tryptophanyl-tRNAs and organization of the corresponding genes", *Journal of Bacteriology*, vol. 174, no. 20, pp. 6471–6478, 1992.

[55] S. Matsuyama, T. Ueda, P. F. Crain, J. A. McCloskey, and K. Watanabe, "A novel wobble rule found in starfish mitochondria. Presence of 7-methylguanosine at the anticodon wobble position expands decoding capability of tRNA", *Journal of Biological Chemistry*, vol. 273, no. 6, pp. 3363–3368, 1998.

[56] K. Tomita, T. Ueda, and K. Watanabe, "7-methylguanosine at the anticodon wobble position of squid mitochondrial $tRNA^{Ser}GCU$: Molecular basis for assignment of AGA/AGG codons as serine in invertebrate mitochondria", *Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression*, vol. 1399, no. 1, pp. 78–82, 1998.

[57] C. Takemoto, L. L. Spremulli, L. A. Benkowski, T. Ueda, T. Yokogawa, and K. Watanabe, " Unconventional decoding of the AUA codon as methionine by mitochondrial tRNA Met with the anticodon f^5CAU as revealed with a mitochondrial in vitro translation system ", *Nucleic Acids Research*, vol. 37, no. 5, pp. 1616–1627, 2009.

[58] T Yokogawa, T Suzuki, T Ueda, *et al.*, "Serine tRNA complementary to the nonuniversal serine codon CUG in *Candida cylindracea*: evolutionary implications", *Proceedings of the National Academy of Sciences*, vol. 89, no. 16, pp. 7408–7411, 1992.

[59] Y. Shulgina and S. R. Eddy, "A computational screen for alternative genetic codes in over 250,000 genomes", *eLife*, vol. 10, e71402, 2021.

[60] R. Giegé and G. Eriani, "The tRNA identity landscape for aminoacylation and beyond", *Nucleic Acids Research*, vol. 51, no. 4, pp. 1528–1570, 2023.

[61] R. Himmelreich, H. Hilbert, H. Plagens, E. Pirkl, B.-C. Li, and R. Herrmann, " Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*", *Nucleic Acids Research*, vol. 24, no. 22, pp. 4420–4449, 1996.

[62] J. P. McCutcheon, B. R. McDonald, and N. A. Moran, "Origin of an alternative genetic code in the extremely small and GC–rich genome of a bacterial symbiont", *PLOS Genetics*, vol. 5, no. 7, pp. 1–11, 2009.

[63] I. Duarte, S. B. Nabuurs, R. Magno, and M. Huynen, "Evolution and diversification of the organellar release factor family", *Molecular Biology and Evolution*, vol. 29, no. 11, pp. 3497–3512, 2012.

[64] J. H. Campbell, P. O'Donoghue, A. G. Campbell, *et al.*, "UGA is an additional glycine codon in uncultured SR1 bacteria from the human microbiota", *Proceedings of the National Academy of Sciences*, vol. 110, no. 14, pp. 5540–5545, 2013.

[65] S. Kervestin, L. Frolova, L. Kisselev, and O. Jean-Jean, "Stop codon recognition in ciliates: *Euplotes* release factor does not respond to reassigned UGA codon", *EMBO reports*, vol. 2, no. 8, pp. 680–684, 2001.

[66] B. D. Eliseev, E. Z. Alkalaeva, P. N. Kryuchkova, *et al.*, "Translation termination factor eRF1 of the ciliate *Blepharisma japonicum* recognizes all three stop codons", *Molecular Biology*, vol. 45, no. 4, pp. 614–618, 2011.

[67] S. Osawa and T. H. Jukes, "Codon reassignment (codon capture) in evolution", *Journal of Molecular Evolution*, vol. 28, no. 4, pp. 271–278, 1989.

[68] S. Sengupta and P. G. Higgs, "A unified model of codon reassignment in alternative genetic codes", *Genetics*, vol. 170, no. 2, pp. 831–840, 2005.

[69] S. Mühlhausen, P. Findeisen, U. Plessmann, H. Urlaub, and M. Kollmar, "A novel nuclear genetic code alteration in yeasts and the evolution of codon reassignment in eukaryotes", *Genome Research*, vol. 26, no. 7, pp. 945–955, 2016.

[70] D. W. Schultz and M. Yarus, "Transfer RNA mutation and the malleability of the genetic code", *Journal of Molecular Biology*, vol. 235, no. 5, pp. 1377–1380, 1994.

[71] T. Maeshiro and M. Kimura, "The role of robustness and changeability on the origin and evolution of genetic codes", *Proceedings of the National Academy of Sciences*, vol. 95, no. 9, pp. 5088–5093, 1998.

[72] P. Błażej, M. Wnętrzak, D. Mackiewicz, P. Gagat, and P. Mackiewicz, "Many alternative and theoretical genetic codes are more robust to amino acid replacements than the standard genetic code", *Journal of Theoretical Biology*, vol. 464, pp. 21–32, 2019.

[73] S. G. E. Andersson and C. G. Kurland, "Genomic evolution drives the evolution of the translation system", *Biochemistry and Cell Biology*, vol. 73, no. 11-12, pp. 775–787, 1995.

[74] S. G. Andersson and C. G. Kurland, "Reductive evolution of resident genomes", *Trends in Microbiology*, vol. 6, no. 7, pp. 263–268, 1998.

[75] L. Gorini, "Informational suppression", *Annual Review of Genetics*, vol. 4, pp. 107–134, 1970.

[76] M. A. S. Santos, C. Cheesman, V. Costa, P. Moradas-Ferreira, and M. F. Tuite, "Selective advantages created by codon ambiguity allowed for the evolution of an alternative genetic code in *Candida* spp.", *Molecular Microbiology*, vol. 31, no. 3, pp. 937–947, 1999.

[77] L. A. Shackelton and E. C. Holmes, "The role of alternative genetic codes in viral evolution and emergence", *Journal of Theoretical Biology*, vol. 254, no. 1, pp. 128–134, 2008.

[78] T. Krassowski, A. Y. Coughlan, X.-X. Shen, *et al.*, "Evolutionary instability of CUG-Leu in the genetic code of budding yeasts", *Nature Communications*, vol. 9, p. 1887, 2018.

[79] P. Alberch, "From genes to phenotype: Dynamical systems and evolvability", *Genetica*, vol. 84, no. 1, pp. 5–11, 1991.

[80] J. Maynard Smith, "Natural selection and the concept of a protein space", *Nature*, vol. 225, no. 5232, pp. 563–564, 1970.

[81] B. M. R. Stadler, P. F. Stadler, G. P. Wagner, and W. Fontana, "The topology of the possible: Formal spaces underlying patterns of evolutionary change", *Journal of Theoretical Biology*, vol. 213, no. 2, pp. 241–274, 2001.

[82] G. H. Gonnet, M. A. Cohen, and S. A. Benner, "Exhaustive matching of the entire protein sequence database", *Science*, vol. 256, no. 5062, pp. 1443–1445, 1992.

[83] J. B. Plotkin, J. Dushoff, and H. B. Fraser, "Detecting selection using a single genome sequence of *M. tuberculosis* and *P. falciparum*", *Nature*, vol. 428, no. 6986, pp. 942–945, 2004.

[84] Y. Chen, J. J. Emerson, and T. M. Martin, "Codon volatility does not detect selection", *Nature*, vol. 433, no. 7023, E6–E7, 2005.

[85] M. W. Hahn, J. G. Mezey, D. J. Begun, *et al.*, "Codon bias and selection on single genomes", *Nature*, vol. 433, no. 7023, E5–E6, 2005.

[86] R. Nielsen and M. J. Hubisz, "Detecting selection needs comparative data", *Nature*, vol. 433, no. 7023, E6, 2005.

[87] J. B. Plotkin, J. Dushoff, and H. B. Fraser, "Codon volatility does not detect selection (reply)", *Nature*, vol. 433, no. 7023, E7–E8, 2005.

[88] U. Hershberg and M. J. Shlomchik, "Differences in potential for amino acid change after mutation reveals distinct strategies for κ and λ light-chain variation", *Proceedings of the National Academy of Sciences*, vol. 103, no. 43, pp. 15 963–15 968, 2006.

[89] G. D. Victora and M. C. Nussenzweig, "Germinal centers", *Annual Review of Immunology*, vol. 30, pp. 429–457, 2012.

[90] C. L. Burch and L. Chao, "Evolvability of an RNA virus is determined by its mutational neighbourhood", *Nature*, vol. 406, no. 6796, pp. 625–628, 2000.

[91] G. Cambray and D. Mazel, "Synonymous genes explore different evolutionary landscapes", *PLOS Genetics*, vol. 4, no. 11, pp. 1–9, 2008.

[92] A. R. Hall, V. F. Griffiths, R. C. MacLean, and N. Colegrave, "Mutational neighbourhood and mutation supply rate constrain adaptation in *Pseudomonas aeruginosa*", *Proceedings of the Royal Society B: Biological Sciences*, vol. 277, no. 1681, pp. 643–650, 2010.

[93] G. Pines, J. D. Winkler, A. Pines, and R. T. Gill, "Refactoring the genetic code for increased evolvability", *mBio*, vol. 8, no. 6, 2017.

[94] W. Zhu and S. Freeland, "The standard genetic code enhances adaptive evolution of proteins", *Journal of Theoretical Biology*, vol. 239, no. 1, pp. 63–70, 2006.

[95] S. Tripathi and M. W. Deem, "The standard genetic code facilitates exploration of the space of functional nucleotide sequences", *Journal of Molecular Evolution*, vol. 86, pp. 325–339, 2018.

[96] S. J. Freeland, T. Wu, and N. Keulmann, "The case for an error minimizing standard genetic code", *Origins of life and evolution of the biosphere*, vol. 33, no. 4, pp. 457–477, 2003.

[97] M. Wnetrzak, P. Błażej, and P. Mackiewicz, "Optimization of the standard genetic code in terms of two mutation types: Point mutations and frameshifts", *Biosystems*, vol. 181, pp. 44–50, 2019.

[98] H. Xu and J. Zhang, "On the origin of frameshift-robustness of the standard genetic code", *Molecular Biology and Evolution*, vol. 38, no. 10, pp. 4301–4309, 2021.

[99] L. Shenhav and D. Zeevi, "Resource conservation manifests in the genetic code", *Science*, vol. 370, no. 6517, pp. 683–687, 2020.

[100] D. Mazel and P. Marlière, "Adaptive eradication of methionine and cysteine from cyanobacterial light-harvesting proteins", *Nature*, vol. 341, no. 6239, pp. 245–248, 1989.

[101] J. J. Elser, W. F. Fagan, S. Subramanian, and S. Kumar, "Signatures of ecological resource availability in the animal and plant proteomes", *Molecular Biology and Evolution*, vol. 23, no. 10, pp. 1946–1951, 2006.

[102] J. G. Bragg and A. Wagner, "Protein carbon content evolves in response to carbon availability and may influence the fate of duplicated genes", *Proceedings of the Royal Society B: Biological Sciences*, vol. 274, no. 1613, pp. 1063–1070, 2007.

[103] J. Lv, N. Li, and D.-K. Niu, "Association between the availability of environmental resources and the atomic composition of organismal proteomes: Evidence from *Prochlorococcus* strains living at different depths", *Biochemical and Biophysical Research Communications*, vol. 375, no. 2, pp. 241–246, 2008.

[104] N. Li, J. Lv, and D.-K. Niu, "Low contents of carbon and nitrogen in highly abundant proteins: Evidence of selection for the economy of atomic composition", *Journal of Molecular Evolution*, vol. 68, no. 3, pp. 248–255, 2009.

[105] J. J. Grzymski and A. M. Dussaq, "The significance of nitrogen cost minimization in proteomes of marine microorganisms", *The ISME Journal*, vol. 6, no. 1, pp. 71–80, 2012.

[106] D. R. Mende, J. A. Bryant, F. O. Aylward, *et al.*, "Environmental drivers of a microbial genomic transition zone in the ocean's interior", *Nature Microbiology*, vol. 2, no. 10, pp. 1367–1373, 2017.

[107] F. L. Hellweger, Y. Huang, and H. Luo, "Carbon limitation drives GC content evolution of a marine bacterium in an individual-based genome-scale model", *The ISME Journal*, vol. 12, no. 5, pp. 1180–1187, 2018.

[108] S. J. Freeland, R. D. Knight, and L. F. Landweber, "Measuring adaptation within the genetic code", *Trends in Biochemical Sciences*, vol. 25, pp. 44–45, 2000.

[109] M. Di Giulio, "The extension reached by the minimization of the polarity distances during the evolution of the genetic code", *Journal of Molecular Evolution*, vol. 29, pp. 288–293, 4 1989.

[110] M. Di Giulio, M. Capobianco, and M. Medugno, "On the optimization of the physicochemical distances between amino acids in the evolution of the genetic code", *Journal of Theoretical Biology*, vol. 168, no. 1, pp. 43–51, 1994.

[111] J. Santos and A. Monteagudo, "Simulated evolution applied to study the genetic code optimality using a model of codon reassignments", *BMC Bioinformatics*, vol. 12, p. 56, 1 2011.

[112] P. Błażej, M. Wnetrzak, D. Mackiewicz, and P. Mackiewicz, "Optimization of the standard genetic code according to three codon positions using an evolutionary algorithm", *PLOS ONE*, vol. 13, no. 8, pp. 1–32, 2018.

[113] M. Wnetrzak, P. Błażej, D. Mackiewicz, and P. Mackiewicz, "The optimality of the standard genetic code assessed by an eight-objective evolutionary algorithm", *BMC Evolutionary Biology*, vol. 18, p. 192, 2018.

[114] C. Alff-Steinberger, "The genetic code and error transmission", *Proceedings of the National Academy of Sciences*, vol. 64, no. 2, pp. 584–591, 1969.

[115] J. G. Caporaso, M. Yarus, and R. D. Knight, "Error minimization and coding triplet/binding site associations are independent features of the canonical genetic code", *Journal of Molecular Evolution*, vol. 61, pp. 597–607, 5 2005.

[116] S. Wichmann and Z. Ardern, "Optimality in the standard genetic code is robust with respect to comparison code sets", *Biosystems*, vol. 185, p. 104 023, 2019.

[117] M. Archetti, "Codon usage bias and mutation constraints reduce the level of error minimization of the genetic code", *Journal of Molecular Evolution*, vol. 59, pp. 258–266, 2 2004.

[118] H. Goodarzi, H. Shateri Najafabadi, and N. Torabi, "On the coevolution of genes and genetic code", *Gene*, vol. 362, pp. 133–140, 2005.

[119] J. Athey, A. Alexaki, E. Osipova, *et al.*, "A new and updated resource for codon usage tables", *BMC Bioinformatics*, vol. 18, no. 1, p. 391, 2017.

[120] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing", *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

[121] H. Rozhoňová and J. L. Payne, "Little evidence the standard genetic code is optimized for resource conservation", *Molecular Biology and Evolution*, vol. 38, no. 11, pp. 5127–5133, 2021.

[122] R. Grantham, "Amino acid difference formula to help explain protein evolution", *Science*, vol. 185, no. 4154, pp. 862–864, 1974.

[123] M. Hasegawa and T. Miyata, "On the antisymmetry of the amino acid code table", *Origins of Life*, vol. 10, no. 3, pp. 265–270, 1980.

[124] M. Di Giulio, "Some aspects of the organization and evolution of the genetic code", *Journal of Molecular Evolution*, vol. 29, pp. 191–201, 3 1989.

[125] M. J. Dufton, "Genetic code synonym quotas and amino acid complexity: Cutting the cost of proteins?", *Journal of Theoretical Biology*, vol. 187, no. 2, pp. 165–173, 1997.

[126] J. M. Ogle, D. E. Brodersen, W. M. Clemons, M. J. Tarry, A. P. Carter, and V. Ramakrishnan, "Recognition of cognate transfer RNA by the 30S ribosomal subunit", *Science*, vol. 292, no. 5518, pp. 897–902, 2001.

[127] J. M. Ogle, A. P. Carter, and V. Ramakrishnan, "Insights into the decoding mechanism from recent ribosome structures", *Trends in Biochemical Sciences*, vol. 28, no. 5, pp. 259–266, 2003.

[128] F. Crick, "Codon—anticodon pairing: The wobble hypothesis", *Journal of Molecular Biology*, vol. 19, no. 2, pp. 548–555, 1966.

[129] A. Stoltzfus and L. Y. Yampolsky, "Amino acid exchangeability and the adaptive code hypothesis", *Journal of Molecular Evolution*, vol. 65, no. 4, pp. 456–462, 2007.

[130] M. Di Giulio, "A non-neutral origin for error minimization in the origin of the genetic code", *Journal of Molecular Evolution*, vol. 86, no. 9, pp. 593–597, 2018.

[131] H. Xu and J. Zhang, "Is the genetic code optimized for resource conservation?", *Molecular Biology and Evolution*, vol. 38, no. 11, pp. 5122–5126, 2021.

[132] S. E. Massey, "A sequential "2-1-3" model of genetic code evolution that explains codon constraints", *Journal of Molecular Evolution*, vol. 62, no. 6, pp. 809–810, 2006.

[133] J. B. Plotkin and G. Kudla, "Synonymous but not the same: The causes and consequences of codon bias", *Nature Reviews Genetics*, vol. 12, no. 1, pp. 32–42, 2011.

[134] T. E. Cole, Y. Hong, C. M. Brasier, and K. W. Buck, "Detection of an RNA-dependent RNA polymerase in mitochondria from a mitovirus-infected isolate of the Dutch Elm disease fungus, *Ophiostoma novo-ulmi*", *Virology*, vol. 268, no. 2, pp. 239–243, 2000.

[135] D. J. Taylor, M. J. Ballinger, S. M. Bowman, and J. Bruenn, "Virus-host co-evolution under a modified nuclear genetic code", *PeerJ*, vol. 1, e50, 2013.

[136] N. N. Ivanova, P. Schwientek, H. J. Tripp, *et al.*, "Stop codon reassignment in the wild", *Science*, vol. 344, no. 6186, pp. 909–913, 2014.

[137] A. E. Devoto, J. M. Santini, M. R. Olm, *et al.*, "Megaphages infect *Prevotella* and variants are widespread in gut microbiomes", *Nature Microbiology*, vol. 4, no. 4, pp. 693–700, 2019.

[138] M. A. Crisci, L.-X. Chen, A. E. Devoto, *et al.*, "Closely related Lak megaphages replicate in the microbiomes of diverse animals", *iScience*, vol. 24, no. 8, p. 102 875, 2021.

[139] N. Yutin, S. Benler, S. A. Shmakov, *et al.*, "Analysis of metagenome-assembled viral genomes from the human gut reveals diverse putative CrAss-like phages with unique genomic features", *Nature Communications*, vol. 12, no. 1, p. 1044, 2021.

[140] A. L. Borges, Y. C. Lou, R. Sachdeva, *et al.*, "Widespread stop-codon recoding in bacteriophages may regulate translation of lytic genes", *Nature Microbiology*, vol. 7, no. 6, pp. 918–927, 2022.

[141] S. L. Peters, A. L. Borges, R. J. Giannone, M. J. Morowitz, J. F. Banfield, and R. L. Hettich, "Experimental validation that human microbiome phages use alternative genetic coding", *Nature Communications*, vol. 13, no. 1, p. 5710, 2022.

[142] A. Pfennig, A. Lomsadze, and M. Borodovsky, "MgCod: Gene prediction in phage genomes with multiple genetic codes", *Journal of Molecular Biology*, vol. 435, no. 14, p. 168 159, 2023.

[143] A. P. Camargo, S. Nayfach, I.-M. A. Chen, *et al.*, "IMG/VR v4: an expanded database of uncultivated virus genomes within a framework of extensive functional, taxonomic, and ecological metadata", *Nucleic Acids Research*, vol. 51, no. D1, pp. D733–D743, 2022.

[144] H. Bin Jang, B. Bolduc, O. Zablocki, *et al.*, "Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks", *Nature Biotechnology*, vol. 37, no. 6, pp. 632–639, 2019.

[145] E. M. Adriaenssens, M. B. Sullivan, P. Knezevic, *et al.*, "Taxonomy of prokaryotic viruses: 2018-2019 update from the ICTV Bacterial and Archaeal Viruses Subcommittee", *Archives of Virology*, vol. 165, no. 5, pp. 1253–1260, 2020.

[146] Y.-M. Hou and P. Schimmel, "A simple structural feature is a major determinant of the identity of a transfer RNA", *Nature*, vol. 333, no. 6169, pp. 140–145, 1988.

[147] W. H. McClain and K. Foss, "Changing the identity of a tRNA by introducing a G-U wobble pair near the 3′ acceptor end", *Science*, vol. 240, no. 4853, pp. 793–796, 1988.

[148] D. Turner, A. N. Shkoporov, C. Lood, *et al.*, "Abolishment of morphology-based taxa and change to binomial species names: 2022 taxonomy update of the ICTV bacterial viruses subcommittee", *Archives of Virology*, vol. 168, no. 2, p. 74, 2023.

[149] H. Asahara, H. Himeno, K. Tamura, N. Nameki, T. Hasegawa, and M. Shimizu, "*Escherichia coli* seryl-tRNA synthetase recognizes $tRNA^{Ser}$ by its characteristics tertiary structure", *Journal of Molecular Biology*, vol. 236, no. 3, pp. 738–748, 1994.

[150] S.-i. Sekine, O. Nureki, K. Sakamoto, *et al.*, "Major identity determinants in the "augmented D helix" of $tRNA^{Glu}$ from *Escherichia coli*", *Journal of Molecular Biology*, vol. 256, no. 4, pp. 685–700, 1996.

[151] J. M. Sherman, K. Rogers, M. J. Rogers, and D. Söll, "Synthetase competition and tRNA context determine the in vivo identity of tRNA discriminator mutants", *Journal of Molecular Biology*, vol. 228, no. 4, pp. 1055–1062, 1992.

[152] P. Chan, B. Lin, A. Mak, and T. Lowe, "tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes", *Nucleic Acids Research*, vol. 49, no. 16, pp. 9077–9096, 2021.

[153] A. Kachale, Z. Pavlíková, A. Nenarokova, *et al.*, "Short tRNA anticodon stem and mutant eRF1 allow stop codon reassignment", *Nature*, vol. 613, no. 7945, pp. 751–758, 2023.

[154] C. Francklyn, J. J. Perona, J. Puetz, and Y.-M. Hou, "Aminoacyl-tRNA synthetases: Versatile players in the changing theater of translation", *RNA*, vol. 8, no. 11, 1363–1372, 2002.

[155] D. Hyatt, G.-L. Chen, P. F. LoCascio, M. L. Land, F. W. Larimer, and L. J. Hauser, "Prodigal: prokaryotic gene recognition and translation initiation site identification", *BMC Bioinformatics*, vol. 11, no. 1, p. 119, 2010.

[156] S. U. Schneider, M. B. Leible, and X.-P. Yang, "Strong homology between the small subunit of ribulose-1,5-bisphosphate carboxylase/oxygenase of two species of *Acetabularia* and the occurrence of unusual codon usage", *Molecular and General Genetics*, vol. 218, no. 3, pp. 445–452, 1989.

[157] E. Cocquyt, G. H. Gile, F. Leliaert, H. Verbruggen, P. J. Keeling, and O. De Clerck, "Complex phylogenetic distribution of a non-canonical genetic code in green algae", *BMC Evolutionary Biology*, vol. 10, no. 1, p. 327, 2010.

[158] Y. Sato, M. Hirayama, K. Morimoto, N. Yamamoto, S. Okuyama, and K. Hori, "High mannose-binding lectin with preference for the cluster of α1–2-mannose from the green alga *Boodlea coacta* is a potent entry inhibitor of HIV-1 and influenza viruses", *Journal of Biological Chemistry*, vol. 286, no. 22, pp. 19 446–19 458, 2011.

[159] F. Caron and E. Meyer, "Does *Paramecium primaurelia* use a different genetic code in its macronucleus?", *Nature*, vol. 314, no. 6007, pp. 185–188, 1985.

[160] E. Helftenbein, "Nucleotide sequence of a macronuclear DNA molecule coding for α-tubulin from the ciliate *Stylonychia lemnae*. Special codon usage: TAA is not a translation termination codon", *Nucleic Acids Research*, vol. 13, no. 2, pp. 415–433, 1985.

[161] S Horowitz and M. A. Gorovsky, "An unusual genetic code in nuclear genes of *Tetrahymena*", *Proceedings of the National Academy of Sciences*, vol. 82, no. 8, pp. 2452–2455, 1985.

[162] N. Hanyu, Y. Kuchino, S. Nishimura, and H. Beier, "Dramatic events in ciliate evolution: alteration of UAA and UAG termination codons to glutamine codons due to anticodon mutations in two *Tetrahymena* $tRNAs^{Gln}$", *The EMBO Journal*, vol. 5, pp. 1307–1311, 1986.

[163] A. Hatanaka, N. Umeda, S. Yamashita, and N. Hirazawa, "Identification and characterization of a putative agglutination/immobilization antigen on the surface of *Cryptocaryon irritans*", *Parasitology*, vol. 134, no. 9, 1163–1174, 2007.

[164] Y. Yan, X. X. Maurer-Alcal'a, R. Knight, S. L. Kosakovsky Pond, and L. A. Katz, "Single-cell transcriptomics reveal a correlation between genome architecture and gene family evolution in Ciliates", *mBio*, vol. 10, no. 6, e02524–19, 2019.

[165] P. J. Keeling and W. F. Doolittle, "A non-canonical genetic code in an early diverging eukaryotic lineage", *The EMBO Journal*, vol. 15, pp. 2285–2290, 1996.

[166] P. J. Keeling and W. F. Doolittle, "Widespread and ancient distribution of a noncanonical genetic code in diplomonads", *Molecular Biology and Evolution*, vol. 14, no. 9, pp. 895–901, 1997.

[167] P. J. Keeling and B. S. Leander, "Characterisation of a non-canonical genetic code in the oxymonad *Streblomastix strix*", *Journal of Molecular Biology*, vol. 326, no. 5, pp. 1337–1349, 2003.

[168] A. P. De Koning, G. P. Noble, A. A. Heiss, J. Wong, and P. J. Keeling, "Environmental PCR survey to determine the distribution of a non-canonical genetic code in uncultivable oxymonads", *Environmental Microbiology*, vol. 10, no. 1, pp. 65–74, 2008.

[169] S. A. Karpov, K. V. Mikhailov, G. S. Mirzaeva, *et al.*, "Obligately phagotrophic aphelids turned out to branch with the earliest-diverging fungi", *Protist*, vol. 164, no. 2, pp. 195–205, 2013.

[170] T. R. Bachvaroff, "A precedented nuclear genetic code with all three termination codons reassigned as sense codons in the syndinean *Amoebophrya* sp. ex *Karlodinium veneficum*", *PLOS ONE*, vol. 14, no. 2, pp. 1–21, 2019.

[171] T. Pánek, D. Žihala, M. Sokol, *et al.*, "Nuclear genetic codes with a different meaning of the UAG and the UAA codon", *BMC Biology*, vol. 15, no. 1, p. 8, 2017.

[172] G. Csárdi and T. Nepusz, "The igraph software package for complex network research", 2006. [Online]. Available: `https://api.semanticscholar.org/CorpusID:16923281`.

[173] W. Li, K. R. O'Neill, D. H. Haft, *et al.*, "RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation", *Nucleic Acids Research*, vol. 49, no. D1, pp. D1020–D1028, 2020.

[174] K. Katoh, K. Misawa, K. Kuma, and T. Miyata, "MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform", *Nucleic Acids Research*, vol. 30, no. 14, pp. 3059–3066, 2002.

[175] K. Katoh, K.-i. Kuma, H. Toh, and T. Miyata, "MAFFT version 5: improvement in accuracy of multiple sequence alignment", *Nucleic Acids Research*, vol. 33, no. 2, pp. 511–518, 2005.

[176] S. Wright, "The roles of mutation, inbreeding, crossbreeding and selection in evolution", *Proceedings of the XI International Congress of Genetics*, vol. 8, pp. 209–222, 1932.

[177] P. A. Romero and F. H. Arnold, "Exploring protein fitness landscapes by directed evolution", *Nature Reviews Molecular Cell Biology*, vol. 10, no. 12, pp. 866–876, 2009.

[178] J. L. Payne and A. Wagner, "The causes of evolvability and their evolution", *Nature Reviews Genetics*, vol. 20, pp. 24–38, 2019.

[179] M. Pigliucci, "Is evolvability evolvable?", *Nature Reviews Genetics*, vol. 9, pp. 75–82, 2008.

[180] S. Kauffman and S. Levin, "Towards a general theory of adaptive walks on rugged landscapes", *Journal of Theoretical Biology*, vol. 128, no. 1, pp. 11–45, 1987.

[181] J. A. G. de Visser and J. Krug, "Empirical fitness landscapes and the predictability of evolution", *Nature Reviews Genetics*, vol. 15, no. 7, pp. 480–490, 2014.

[182] R. D. Knight, S. J. Freeland, and L. F. Landweber, "Selection, history and chemistry: The three faces of the genetic code", *Trends in Biochemical Sciences*, vol. 24, no. 6, pp. 241–247, 1999.

[183] E. V. Koonin and A. S. Novozhilov, "Origin and evolution of the genetic code: The universal enigma", *IUBMB Life*, vol. 61, no. 2, pp. 99–111, 2009.

[184] R. A. Fisher, *A Genetical Theory of Natural Selection*. Oxford: Clarendon Press, 1930.

[185] S. J. Freeland, "The Darwinian genetic code: An adaptation for adapting?", *Genetic Programming and Evolvable Machines*, vol. 3, no. 2, pp. 113–127, 2002.

[186] E. Firnberg, J. W. Labonte, J. J. Gray, and M. Ostermeier, "A comprehensive, high-resolution map of a gene's fitness landscape", *Molecular Biology and Evolution*, vol. 31, no. 6, pp. 1581–1592, 2014.

[187] N. Ostrov, M. Landon, M. Guell, *et al.*, "Design, synthesis, and testing toward a 57-codon genome", *Science*, vol. 353, no. 6301, pp. 819–822, 2016.

[188] D. de la Torre and J. W. Chin, "Reprogramming the genetic code", *Nature Reviews Genetics*, vol. 22, no. 3, pp. 169–184, 2021.

[189] J. Fredens, K. Wang, D. de la Torre, *et al.*, "Total synthesis of *Escherichia coli* with a recoded genome", *Nature*, vol. 569, no. 7757, pp. 514–518, 2019.

[190] T. Aita, S. Urata, and Y. Husimi, "From amino acid landscape to protein landscape: Analysis of genetic codes in terms of fitness landscape", *Journal of Molecular Evolution*, vol. 50, pp. 313–323, 2000.

[191] M. Karageorgi, S. C. Groen, F. Sumbul, *et al.*, "Genome editing retraces the evolution of toxin resistance in the monarch butterfly", *Nature*, vol. 574, no. 7778, pp. 409–412, 2019.

[192] C. Natarajan, A. Jendroszek, A. Kumar, *et al.*, "Molecular basis of hemoglobin adaptation in the high-flying bar-headed goose", *PLOS Genetics*, vol. 14, no. 4, pp. 1–19, 2018.

[193] R. Fasan, Y. T. Meharenna, C. D. Snow, T. L. Poulos, and F. H. Arnold, "Evolutionary history of a specialized P450 propane monooxygenase", *Journal of Molecular Biology*, vol. 383, no. 5, pp. 1069–1080, 2008.

[194] M. Goldsmith and D. S. Tawfik, "Enzyme engineering: Reaching the maximal catalytic efficiency peak", *Current Opinion in Structural Biology*, vol. 47, pp. 140–150, 2017.

[195] J. Calles, I. Justice, D. Brinkley, A. Garcia, and D. Endy, "Fail-safe genetic codes designed to intrinsically contain engineered organisms", *Nucleic Acids Research*, vol. 47, no. 19, pp. 10 439–10 451, 2019.

[196] J. B. Kinney and D. M. McCandlish, "Massively parallel assays and quantitative sequence–function relationships", *Annual Review of Genomics and Human Genetics*, vol. 20, no. 1, pp. 99–127, 2019.

[197] N. C. Wu, L. Dai, C. A. Olson, J. O. Lloyd-Smith, and R. Sun, "Adaptation in protein fitness landscapes is facilitated by indirect paths", *eLife*, vol. 5, e16965, 2016.

[198] T.-L. V. Lite, R. A. Grant, I. Nocedal, M. L. Littlehale, M. S. Guo, and M. T. Laub, "Uncovering the basis of protein-protein interaction specificity with a combinatorially complete library", *eLife*, vol. 9, e60924, 2020.

[199] E. C. Hartman, M. J. Lobba, A. H. Favor, S. A. Robinson, M. B. Francis, and D. Tullman-Ercek, "Experimental evaluation of coevolution in a self-assembling particle", *Biochemistry*, vol. 58, no. 11, pp. 1527–1538, 2019.

[200] A. S. Jalal, N. T. Tran, C. E. Stevenson, *et al.*, "Diversification of DNA-binding specificity by permissive and specificity-switching mutations in the ParB/Noc protein family", *Cell Reports*, vol. 32, no. 3, 2020.

[201] A. Papkou, L. Garcia-Pastor, J. A. Escudero, and A. Wagner, "A rugged yet easily navigable fitness landscape", *Science*, vol. 382, no. 6673, eadh3860, 2023.

[202] U Sjöbring, L Björck, and W Kastern, "Streptococcal protein G. Gene structure and protein binding properties.", *Journal of Biological Chemistry*, vol. 266, no. 1, pp. 399–405, 1991.

[203] A. Sauer-Eriksson, G. J. Kleywegt, M. Uhlén, and T. Jones, "Crystal structure of the C2 fragment of streptococcal protein G in complex with the Fc domain of human IgG", *Structure*, vol. 3, no. 3, pp. 265–278, 1995.

[204] C. A. Olson, N. C. Wu, and R. Sun, "A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain", *Current Biology*, vol. 24, no. 22, pp. 2643–2651, 2014.

[205] N. Fraikin, F. Goormaghtigh, and L. V. Melderen, "Type II toxin-antitoxin systems: Evolution and revolutions", *Journal of Bacteriology*, vol. 202, no. 7, e00763–19, 2020.

[206] D. C.-H. Lin and A. D. Grossman, "Identification and characterization of a bacterial chromosome partitioning site", *Cell*, vol. 92, no. 5, pp. 675–685, 1998.

[207] E. Toprak, A. Veres, J.-B. Michel, R. Chait, D. L. Hartl, and R. Kishony, "Evolutionary paths to antibiotic resistance under dynamically sustained drug selection", *Nature Genetics*, vol. 44, no. 1, pp. 101–105, 2012.

[208] Y. T. Tamer, I. K. Gaszek, H. Abdizadeh, *et al.*, "High-order epistasis in catalytic power of dihydrofolate reductase gives rise to a rugged fitness landscape in the presence of trimethoprim selection", *Molecular Biology and Evolution*, vol. 36, no. 7, pp. 1533–1550, 2019.

[209] N. Tokuriki and D. S. Tawfik, "Protein dynamism and evolvability", *Science*, vol. 324, no. 5924, pp. 203–207, 2009.

[210] J. Zhou, M. S. Wong, W.-C. Chen, A. R. Krainer, J. B. Kinney, and D. M. McCandlish, "Higher-order epistasis and phenotypic prediction", *Proceedings of the National Academy of Sciences*, vol. 119, no. 39, e2204233119, 2022.

[211] A. Wagner, "Neutralism and selectionism: A network-based reconciliation", *Nature Reviews Genetics*, vol. 9, pp. 965–974, 2009.

[212] J. Aguilar-Rodríguez, J. L. Payne, and A. Wagner, "A thousand empirical adaptive landscapes and their navigability", *Nature Ecology & Evolution*, vol. 1, no. 2, p. 0045, 2017.

[213] I. G. Szendro, M. F. Schenk, J. Franke, J. Krug, and J. A. G. M. de Visser, "Quantitative analyses of empirical fitness landscapes", *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2013, no. 01, P01005, 2013.

[214] D. M. Weinreich, R. A. Watson, and L. Chao, "Perspective: Sign epistasis and genetic costraint on evolutionary trajectories", *Evolution*, vol. 59, no. 6, pp. 1165–1174, 2005.

[215] F. J. Poelwijk, D. J. Kiviet, D. M. Weinreich, and S. J. Tans, "Empirical fitness landscapes reveal accessible evolutionary paths", *Nature*, vol. 445, no. 7126, pp. 383–386, 2007.

[216] D. M. Weinreich, N. F. Delaney, M. A. DePristo, and D. L. Hartl, "Darwinian evolution can follow only very few mutational paths to fitter proteins", *Science*, vol. 312, no. 5770, pp. 111–114, 2006.

[217] J. Franke, A. Klözer, J. A. G. M. de Visser, and J. Krug, "Evolutionary accessibility of mutational pathways", *PLOS Computational Biology*, vol. 7, no. 8, pp. 1–9, 2011.

[218] S. Kawashima, H. Ogata, and M. Kanehisa, "AAindex: Amino Acid Index Database", *Nucleic Acids Research*, vol. 27, no. 1, pp. 368–369, 1999.

[219] S. Kawashima and M. Kanehisa, "AAindex: Amino Acid index database", *Nucleic Acids Research*, vol. 28, no. 1, p. 374, 2000.

[220] D. M. McCandlish, "Visualizing fitness landscapes", *Evolution*, vol. 65, no. 6, pp. 1544–1558, 2011.

[221] J. Zhou and D. M. McCandlish, "Minimum epistasis interpolation for sequence-function relationships", *Nature Communications*, vol. 11, no. 1, p. 1782, 2020.

[222] T. Mukai, M. J. Lajoie, M. Englert, and D. Söll, "Rewriting the genetic code", *Annual Review of Microbiology*, vol. 71, no. 1, pp. 557–577, 2017.

[223] J. W. Chin, T. A. Cropp, J. C. Anderson, M. Mukherji, Z. Zhang, and P. G. Schultz, "An expanded eukaryotic genetic code", *Science*, vol. 301, no. 5635, pp. 964–967, 2003.

[224] H. Dong, L. Nilsson, and C. G. Kurland, "Co-variation of tRNA abundance and codon usage in *Escherichia coli* at different growth rates", *Journal of Molecular Biology*, vol. 260, no. 5, pp. 649–663, 1996.

[225] P. F. Agris, E. R. Eruysal, A. Narendran, V. Y. P. Väre, S. Vangaveti, and S. V. Ranganathan, "Celebrating wobble decoding: Half a century and still much is new", *RNA Biology*, vol. 15, no. 4-5, pp. 537–553, 2018.

[226] A. Wagner, *Robustness and Evolvability in Living Systems*. Princeton University Press, 2005.

[227] M. Lässig, V. Mustonen, and A. Walczak, "Predicting evolution", *Nature Ecology and Evolution*, vol. 1, 2017.

[228] D. M. McCandlish, "On the findability of genotypes", *Evolution*, vol. 67, no. 9, pp. 2592–2603, 2013.

[229] K. Dingle, F. Ghaddar, P. Šulc, and A. A. Louis, "Phenotype bias determines how natural rna structures occupy the morphospace of all possible shapes", *Molecular Biology and Evolution*, vol. 39, no. 1, 2021.

[230] S. Schaper and A. A. Louis, "The arrival of the frequent: How bias in genotype-phenotype maps can steer populations to local optima", *PLOS ONE*, vol. 9, no. 2, pp. 1–9, 2014.

[231] J. L. King and T. H. Jukes, "Non-Darwinian evolution", *Science*, vol. 164, no. 3881, pp. 788–798, 1969.

[232] J. L. King, "The role of mutation in evolution", *Sixth Berkeley Symposium on Mathematical Statistics and Probability*, 1973.

[233] S. Gavrilets, "High-Dimensional Fitness Landscapes and Speciation", in *Evolution—the Extended Synthesis*, The MIT Press, 2010.

[234] S. F. Greenbury, A. A. Louis, and S. E. Ahnert, "The structure of genotype-phenotype maps makes fitness landscapes navigable", *Nature Ecology & Evolution*, vol. 6, no. 11, pp. 1742–1752, 2022.

[235] M. Dayhoff, R. Schwartz, and B. Orcutt, "A model of evolutionary change in proteins", in *Atlas of Protein Sequence and Structure*, M. Dayhoff, Ed., vol. 5, Washington, D. C.: National Biomedical Research Foundation, 1978, pp. 345–352.

[236] S Henikoff and J. G. Henikoff, "Amino acid substitution matrices from protein blocks", *Proceedings of the National Academy of Sciences*, vol. 89, no. 22, pp. 10 915–10 919, 1992.

[237] L. Y. Yampolsky and A. Stoltzfus, "The exchangeability of amino acids in proteins", *Genetics*, vol. 170, no. 4, pp. 1459–1472, 2005.

[238] A. Radványi and A. Kun, "Phylogenetic analysis of mutational robustness based on codon usage supports that the standard genetic code does not prefer extreme environments", *Scientific Reports*, vol. 11, no. 1, p. 10 963, 2021.

[239] A. V. Cano and J. L. Payne, "Mutation bias interacts with composition bias to influence adaptive evolution", *PLOS Computational Biology*, vol. 16, no. 9, pp. 1–26, 2020.

[240] X. Li and C. C. Liu, "Biological applications of expanded genetic codes", *ChemBioChem*, vol. 15, no. 16, pp. 2335–2341, 2014.

[241] X. Jin, O.-J. Park, and S. H. Hong, "Incorporation of non-standard amino acids into proteins: Challenges, recent achievements, and emerging applications", *Applied Microbiology and Biotechnology*, vol. 103, no. 7, pp. 2947–2958, 2019.

[242] P. Marliere, "The farther, the safer: A manifesto for securely navigating synthetic species away from the old living world", *Systems and Synthetic Biology*, vol. 3, no. 1, pp. 77–84, 2009.

[243] V. Kubyshkin and N. Budisa, "Synthetic alienation of microbial organisms by using genetic code engineering: Why and how?", *Biotechnology Journal*, vol. 12, no. 8, p. 1 600 097, 2017.

[244] T. Fujino, M. Tozaki, and H. Murakami, "An amino acid-swapped genetic code", *ACS Synthetic Biology*, vol. 9, no. 10, pp. 2703–2713, 2020.

[245] F. E. Romesberg, "Discovery, implications and initial use of semi-synthetic organisms with an expanded genetic alphabet/code", *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 378, no. 1871, p. 20 220 030, 2023.

[246] S. B. Sun, P. G. Schultz, and C. H. Kim, "Therapeutic applications of an expanded genetic code", *ChemBioChem*, vol. 15, no. 12, pp. 1721–1729, 2014.

[247] J. L. Ptacin, C. E. Caffaro, L. Ma, *et al.*, "An engineered IL-2 reprogrammed for anti-tumor therapy using a semi-synthetic organism", *Nature Communications*, vol. 12, 1 2021.

[248] C. A. L. McFeely, B. Shakya, C. A. Makovsky, A. K. Haney, T. Ashton Cropp, and M. C. T. Hartman, "Extensive breaking of genetic code degeneracy with non-canonical amino acids", *Nature Communications*, vol. 14, no. 1, p. 5008, 2023.

[249] M. J. Lajoie, A. J. Rovner, D. B. Goodman, *et al.*, "Genomically recoded organisms expand biological functions", *Science*, vol. 342, no. 6156, pp. 357–360, 2013.

[250] Y. H. Lau, F. Stirling, J. Kuo, *et al.*, "Large-scale recoding of a bacterial genome by iterative recombineering of synthetic DNA", *Nucleic Acids Research*, vol. 45, no. 11, pp. 6971–6980, 2017.

[251] K. Wang, J. Fredens, S. F. Brunner, S. H. Kim, T. Chia, and J. W. Chin, "Defining synonymous codon compression schemes by genome recoding", *Nature*, vol. 539, no. 7627, pp. 59–64, 2016.

[252] M. J. Hammerling, J. W. Ellefson, D. R. Boutz, E. M. Marcotte, A. D. Ellington, and J. E. Barrick, "Bacteriophages use an expanded genetic code on evolutionary paths to higher fitness", *Nature Chemical Biology*, vol. 10, no. 3, pp. 178–180, 2014.

[253] M. J. Hammerling, J. Gollihar, C. Mortensen, R. N. Alnahhas, A. D. Ellington, and J. E. Barrick, "Expanded genetic codes create new mutational routes to rifampicin resistance in *Escherichia coli*", *Molecular Biology and Evolution*, vol. 33, no. 8, pp. 2054–2063, 2016.

[254] D. S. Tack, A. C. Cole, R. Shroff, B. R. Morrow, and A. D. Ellington, "Evolving bacterial fitness with an expanded genetic code", *Scientific Reports*, vol. 8, no. 1, p. 3288, 2018.

[255] R. Thyer, R. Shroff, D. R. Klein, *et al.*, "Custom selenoprotein production enabled by laboratory evolution of recoded bacterial strains", *Nature Biotechnology*, vol. 36, no. 7, pp. 624–631, 2018.

[256] M. Srivastava, H. Rozhoňová, and J. L. Payne, "Alphabet cardinality and adaptive evolution", *Journal of Physics A: Mathematical and Theoretical*, vol. 56, no. 45, p. 455 601, 2023.

[257] A. F. Rubin, H. Gelman, N. Lucas, *et al.*, "A statistical framework for analyzing deep mutational scanning data", *Genome Biology*, vol. 18, no. 1, p. 150, 2017.

[258] D. Aldous and J. A. Fill, *Reversible Markov Chains and Random Walks on Graphs*. 2002, Unfinished monograph, recompiled 2014, available at `http://www.stat.berkeley.edu/users/aldous/RWG/book.html`.

[259] J. Norris, *Markov Chains*. 1997, ISBN: 978-0-521-48181-6.

[260] A. Nisthal, C. Y. Wang, M. L. Ary, and S. L. Mayo, "Protein stability engineering insights revealed by domain-wide comprehensive mutagenesis", *Proceedings of the National Academy of Sciences*, vol. 116, no. 33, pp. 16 367–16 377, 2019.

[261] S. Kakraba and D. Knisley, "A graph-theoretic model of single point mutations in the cystic fibrosis transmembrane conductance regulator", *Journal of Advances in Biotechnology*, vol. 6, no. 1, 780–786, 2016.

[262] D. M. Dawson, in *The biochemical genetics of man*, D. Brock and O. Mayo, Eds., New York: Academic Press, 1972, pp. 1–38.

[263] M. Levitt, "A simplified representation of protein conformations for rapid simulation of protein folding", *Journal of Molecular Biology*, vol. 104, no. 1, pp. 59–107, 1976.

[264] P. Koehl and M. Levitt, "Structure-based conformational preferences of amino acids", *Proceedings of the National Academy of Sciences*, vol. 96, no. 22, pp. 12 524–12 529, 1999.

[265] F. R. Maxfield and H. A. Scheraga, "Status of empirical methods for the prediction of protein backbone topography", *Biochemistry*, vol. 15, no. 23, pp. 5138–5153, 1976.

[266] P. Sneath, "Relations between chemical structure and biological activity in peptides", *Journal of Theoretical Biology*, vol. 12, no. 2, pp. 157–195, 1966.

[267] K Yutani, K Ogasahara, T Tsujita, and Y Sugino, "Dependence of conformational stability on hydrophobicity of the amino acid residue in a series of variant proteins substituted at a unique position of tryptophan synthase alpha subunit", *Proceedings of the National Academy of Sciences*, vol. 84, no. 13, pp. 4441–4444, 1987.

[268] N. Qian and T. J. Sejnowski, "Predicting the secondary structure of globular proteins using neural network models", *Journal of Molecular Biology*, vol. 202, no. 4, pp. 865–884, 1988.

[269] E. J. Cohn and J. T. Edsall, *Protein, Amino Acid, and Peptides*. New York: Reinhold, 1943.

[270] M. Levitt, "Conformational preferences of amino acids in globular proteins", *Biochemistry*, vol. 17, no. 20, pp. 4277–4285, 1978.

[271] S. Tanaka and H. A. Scheraga, "Statistical mechanical treatment of protein conformation. 5. Multistate model for specific-sequence copolymers of amino acids", *Macromolecules*, vol. 10, no. 1, pp. 9–20, 1977.

[272] H. Nakashima, K. Nishikawa, and T. Ooi, "Distinct character in hydrophobicity of amino acid compositions of mitochondrial proteins", *Proteins: Structure, Function, and Bioinformatics*, vol. 8, no. 2, pp. 173–178, 1990.

[273] J.-L. Fauchère, M. Charton, L. B. Kier, A. Verloop, and V. Pliska, "Amino acid side chain parameters for correlation studies in biology and pharmacology", *International Journal of Peptide and Protein Research*, vol. 32, no. 4, pp. 269–278, 1988.

[274] H. Nakashima and K. Nishikawa, "The amino acid composition is different between the cytoplasmic and extracellular sides in membrane proteins", *FEBS Letters*, vol. 303, no. 2-3, pp. 141–146, 1992.

[275] A. V. Finkelstein, A. Y. Badretdinov, and O. B. Ptitsyn, "Physical reasons for secondary structure stability: α-helices in short peptides", *Proteins: Structure, Function, and Bioinformatics*, vol. 10, no. 4, pp. 287–299, 1991.

[276] V. Muñoz and L. Serrano, "Intrinsic secondary structure propensities of the amino acids, using statistical ϕ–ψ matrices: Comparison with experimental scales", *Proteins: Structure, Function, and Bioinformatics*, vol. 20, no. 4, pp. 301–311, 1994.

[277] B. Robson and E. Suzuki, "Conformational properties of amino acid residues in globular proteins", *Journal of Molecular Biology*, vol. 107, no. 3, pp. 327–356, 1976.

[278] M. Blaber, X. Zhang, and B. W. Matthews, "Structural basis of amino acid α helix propensity", *Science*, vol. 260, no. 5114, pp. 1637–1640, 1993.

[279] D. Eisenberg and A. D. McLachlan, "Solvation energy in protein folding and binding", *Nature*, vol. 319, no. 6050, pp. 199–203, 1986.

[280] R. E. Jacobs and S. H. White, "The nature of the hydrophobic binding of small peptides at the bilayer interface: Implications for the insertion of transbilayer helices", *Biochemistry*, vol. 28, no. 8, pp. 3421–3437, 1989.

[281] D. D. Jones, "Amino acid properties and side-chain orientation in proteins: A cross correlation approach", *Journal of Theoretical Biology*, vol. 50, no. 1, pp. 167–183, 1975.

[282] J. O. Hutchens, "Heat capacities, absolute entropies, and entropies of formation of amino acids and related compounds", in *Handbook of Biochemistry*, H. A. Sober, Ed., Cleveland, Ohio: Chemical Rubber Co., 1970, B60–B61.

[283] G. Khanarian and W. J. Moore, "The Kerr effect of amino acids in water", *Australian Journal of Chemistry*, vol. 33, pp. 1727–1741, 1980.

[284] A. Radzicka and R. Wolfenden, "Comparing the polarities of the amino acids: Side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution", *Biochemistry*, vol. 27, no. 5, pp. 1664–1670, 1988.

[285] H. Guy, "Amino acid side-chain partition energies and distribution of residues in soluble proteins", *Biophysical Journal*, vol. 47, no. 1, pp. 61–70, 1985.

[286] J. S. Richardson and D. C. Richardson, "Amino acid preferences for specific locations at the ends of α helices", *Science*, vol. 240, no. 4859, pp. 1648–1652, 1988.

[287] A. Hopfinger, *Intermolecular interactions and biomolecular organizations*. New York: Wiley, 1977.

[288] C. R. Woese, "Evolution of the genetic code", *Naturwissenschaften*, vol. 60, no. 10, pp. 447–459, 1973.

[289] P. Y. Chou and G. D. Fasman, "Prediction of the secondary structure of proteins from their amino acid sequence", in *Advances in Enzymology and Related Areas of Molecular Biology*. John Wiley & Sons, Ltd, 1979, pp. 45–148.

[290] P. Klein, M. Kanehisa, and C. DeLisi, "Prediction of protein function from sequence properties: Discriminant analysis of a data base", *Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology*, vol. 787, no. 3, pp. 221–226, 1984.

[291] R. Aurora and G. D. Rosee, "Helix capping", *Protein Science*, vol. 7, no. 1, pp. 21–38, 1998.

[292] P. A. P. Moran, "Random processes in genetics", *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 54, no. 1, 60–71, 1958.

[293] A. Stoltzfus and L. Y. Yampolsky, "Climbing Mount Probable: Mutation as a cause of nonrandomness in evolution", *Journal of Heredity*, vol. 100, no. 5, pp. 637–647, 2009.

[294] A. V. Cano, H. Rozhoňová, A. Stoltzfus, D. M. McCandlish, and J. L. Payne, "Mutation bias shapes the spectrum of adaptive substitutions", *Proceedings of the National Academy of Sciences*, vol. 119, no. 7, e2119720119, 2022.

ACKNOWLEDGEMENTS

Many individuals and institutions were instrumental during the course of my PhD and I would like to take this opportunity to thank them.

First and foremost, Josh, you were a wonderful advisor. You gave me the freedom to explore what I found most interesting, yet you were always available for discussions when needed. Your enthusiasm for science is truly infectious. I also appreciate your unwavering support for my major life decisions over the past four years, whether it was expanding my family or moving across the ocean for a year-long research stay at Harvard.

Speaking of my time at Harvard, it would not have been possible without the support of the SNSF Mobility grant. Many people assisted with administrative matters on both sides of the Atlantic, but special thanks go to Rita and Ariane for always being there to address my questions. My deepest gratitude goes to Sean for becoming my advisor for the year, and for guiding me with patience, encouragement, and expertise throughout the research in Chapter 3.

The past four years have been much brighter thanks to the wonderful people I met at both IBZ and Harvard. Alejandro, Alex, Magda, Malvika, and Paco, we may not have been the most productive team, but we sure had a lot of fun together. Thank you, Andreas, Lisa, and the other inhabitants of the H74 office, for keeping me company and making the last few months enjoyable, and thank you, all the other members of the Bonhoeffer lab, for the many interesting conversations over lunch. To all the members of the Eddy lab, thank you for welcoming me so warmly to Cambridge, for bearing with my rants about all the things that are better in Europe than in the US, and for introducing me to the world of crosswords. I miss you all.

I also wish to thank Pedro Beltrao and Stephen Freeland for serving as members of my doctoral committee and Sebastian Bonhoeffer for being my official advisor in the last year. Berit Siedentop and Alejandro Cano shared their thesis templates with me, and Chris Witzany kindly read the automatically generated Zusammenfassung of this thesis and translated it into proper German. Thank you all!

Finally, throughout this thesis, I have drawn parallels between the genetic code and language. Yet, language is not enough to express the love I feel for the two most important people in my life. Vašík and Toníček, thank you for all the joy you bring to my life.

I also wish to acknowledge the assistance of ChatGPT-4 in writing this thesis. The AI's input was used solely for correcting grammar and improving language.