biophysica

MDPI

Article

Genotype-to-Protein Map and Collective Adaptation in a Viral Population

Ariadna Villanueva [1], Henry Secaira-Morocho [1], Luis F. Seoane [1,2], Ester Lázaro [3] and Susanna Manrubia [1,2,*]

[1] National Centre for Biotechnology (CNB-CSIC), c/Darwin 3, Campus de Cantoblanco (UAM), 28049 Madrid, Spain; ariadna.villanuevam@alumnos.upm.es (A.V.); henrry.secairamorocho@gmail.com (H.S.-M.); lf.seoane@cnb.csic.es (L.F.S.)
[2] Grupo Interdisciplinar de Sistemas Complejos (GICS), 28049 Madrid, Spain
[3] Centro de Astrobiología (CAB), INTA-CSIC, Ctra de Ajalvir Km 4, 28850 Madrid, Spain; lazarole@cab.inta-csic.es
* Correspondence: smanrubia@cnb.csic.es

Abstract: Viral populations are large and highly heterogeneous. Despite the evolutionary relevance of such heterogeneity, statistical approaches to quantifying the extent to which viruses maintain a high genotypic and/or phenotypic diversity have been rarely pursued. Here, we address this issue by analyzing a nucleotide-to-protein sequence map through deep sequencing of populations of the Qβ phage adapted to high temperatures. Tens of thousands of different sequences corresponding to two fragments of the gene coding for the viral replicase were recovered. A diversity analysis of two independent populations consistently revealed that about 40% of the mutations identified caused changes in protein amino acids, leading to an almost complete exploration of the protein neighborhood of (non-silent) mutants at a distance of one. The functional form of the empirical distribution of phenotype abundance agreed with analytical calculations that assumed random mutations in the nucleotide sequence. Our results concur with the idea that viral populations maintain a high diversity as an efficient adaptive mechanism and support the hypothesis of universality for a lognormal distribution of phenotype abundances in biologically meaningful genotype–phenotype maps, highlighting the relevance of entropic effects in molecular evolution.

Keywords: genotype–phenotype map; viral quasispecies; deep sequencing; phenotypic bias; non-silent mutations; in vitro evolution; genotype networks

Citation: Villanueva, A.; Secaira-Morocho, H.; Seoane, L.F.; Lázaro, E.; Manrubia, S. Genotype-to-Protein Map and Collective Adaptation in a Viral Population. *Biophysica* **2022**, *2*, 381–399. https://doi.org/10.3390/biophysica2040034

Academic Editor: Jaume Casademunt

Received: 28 September 2022
Accepted: 14 October 2022
Published: 27 October 2022

1. Introduction

Life is endowed with an amazing ability to adapt to harsh circumstances. This ability is particularly prominent at the smallest scales of definition of organisms, where sequences of nucleotides (RNA or DNA, in most cases) change relentlessly to produce raw variations that subsequently modify the phenotypes of organisms, where natural selection eventually acts. Knowledge of the breath of variation in genotypes (molecular sequences of nucleotides in the current context) that extant organisms hold has significantly changed in the last 80 years. In the mid-1930s, mutations were akin to new alleles, whose fate was either to become fixed in the population or disappear. It was broadly assumed that variants had either a positive or a negative effect on the phenotype and rarely appeared in populations, which were, therefore, depicted as homogeneous ensembles of organisms.

Over a decade had to elapse before the nature of the hereditary substrate was uncovered and began to be understood: DNA was shown to be the genetic material of bacteria in the mid-1940s; the structure of the DNA molecule was described in 1953; the first sequence of RNA was obtained in 1965. The first efficient sequencing technique dates back to 1977. In turn, however, formal descriptions of the dynamics of population genetics largely relied on original assumptions and mostly conformed to simple deterministic rules [1], with

exceptions such as Moran's or Wright–Fisher's models (reviewed in [2]). The separation of time scales between the generation of variation and its fixation in a population kept aside the stochastic nature of evolutionary processes behind the generation of diversity and its consequences.

Conceptually, a remarkable breakthrough arrived when Motoo Kimura, upon examining DNA sequencing evidence, realized that mutations were actually abundant in coding sequences, but most of them were neutral, with no effects on protein sequences and, therefore, on phenotypes [3,4]. Contrary to previous assumptions, this observation suggested that populations could bear variation at no cost and accumulate it through genetic drift. The clash between scientific communities adhering to the one-gene–one-function paradigm and those embracing the possibility of a large molecular diversity within populations is well reflected in two papers signed by Frank B. Salisbury [5] and John Maynard Smith [6]. The colossal size of genotype spaces, Salisbury argued, absolutely prevented the achievement of the smallest functional molecule. To this statement, Maynard Smith replied with his fine intuition that, instead, the resilience of molecular function supported the existence of *protein networks*, which must form *"a continuous network which can be traversed by unit mutational steps without passing through nonfunctional intermediates"* [6].

Fifty years after Maynard Smith's paper, no doubts remain regarding the high genotype-to-phenotype (GP) redundancy that underlies molecular function [7]. Computational and empirical studies have shown that the same molecular phenotype can be obtained from an astronomically large number of different genotypes [8] and that such genotypes indeed allow the cost-less exploration of genotype spaces. It is in this way that molecular innovation, in the form of new, accessible phenotypes, occurs [9].

The set of genotypes that map to a phenotype can be typically depicted as a large *genotype network* that spans the whole space of genotypes. That is, almost any two genotypes in that set are mutually accessible through a sequence of mutations, thus guaranteeing the navigability of genotype spaces. The RNA-sequence-to-secondary-structure (S3) map has been, since the early 1990s, an inspiring model system for probing the architecture of the GP map. Many different properties of the latter were first described in an RNA S3 map and, subsequently, shown to be shared by a large number of biologically sensible GP maps [8,10,11]. Some examples of such properties are very high redundancy [12] and a skewed distribution of phenotype abundance or phenotypic bias [13–16], the increase in robustness with the size of the phenotype [17,18], or the preference of highly neutral sequences for clustering together (and vice versa) [18], a property known as assortativity in the parlance of complex networks. Our current understanding of the GP map depicts genotype spaces with a network-of-networks structure in a very high-dimensional space [19] such that, even if random searches are necessarily local, most abundant phenotypes can be found within a small radius (in terms of the number of mutations) from any initial sequence; this is the idea of shape space covering, as first introduced for RNA S3 [20]. The overall structure of genotype spaces has consequences in the evolutionary process [21], such as the unavoidable appearance of punctuated dynamics at various levels [22,23].

The subset of genome spaces that a natural molecular population visits depends on multiple variables—notably, on the population size, the mutation rate, and the time allowed for exploration [24]. Larger populations and higher mutation rates, as well as shorter generation times, should better cover the unfathomably large region available for exploration along adaptation. These features are broadly accepted to be adaptive traits that are responsible for the evolutionary success of viruses—especially those with RNA genomes, with their typically higher mutation rate [25], are suitable systems for comparing theoretical expectations with natural observations. As a practical example of their plasticity, SARS-CoV-2 dramatically illustrates how a steady generation of variation translates, in a relatively short time span, into the emergence of new phenotypes that escape selective pressures and guarantee survival [26]. Still, the collective movement of viral populations in genome spaces remains largely unknown. This study delves into the diversity of viruses, an important milestone towards understanding how molecular

populations explore Maynard Smith's networks and find new phenotypes that allow persistence under changing environmental conditions.

Specifically, here, we quantify the standing genotypic and phenotypic diversity of Qβ as a representative of a class of systems with a large population size in a high-mutation regime. The Qβ phage was the first virus in which high heterogeneity was described, leading to the conclusion that two genomes in a viral population of the phage differed in about one or two nucleotide positions [27]. Subsequent studies with Qβ and other viruses have shown that high diversity is a constitutive characteristic of such organisms [28] and is central for endowing them with a high adaptive ability. Indeed, high adaptability has to be a trait characteristic of any viral species, since, perhaps with rare (unknown, as of yet) exceptions, viruses in their multiple forms infect all cellular organisms on Earth. Previous works on the Qβ phage have already illustrated its adaptive ability under various conditions [29–32] and have characterized some of the main mutations driving adaptation. Recent experiments [24] on adaptation of the Qβ phage have revealed multiple non-silent mutations in its sequence that reach high abundance (above 50%) in different populations, thus appearing in the consensus sequence of populations, as characterized through Sanger sequencing [33]. This variability is indirectly indicative of the possible existence of multiple adaptive pathways—that is, of various mutations occurring in the neighborhood of dominating variants that are able to yield similar adaptive advantages.

In the current study, we first explored nine independent mutations that changed the protein sequence and have been repeatedly observed in experiments on the adaptation of the Qβ phage to high temperatures. This analysis illustrates the difficulty in determining the effects of non-neutral mutations in the phenotype and their possible dependence on other mutations in the population. Subsequently, we analyzed the diversity of two regions of the phage genome in two different populations adapted to a temperature of 43 °C. Each population resulted from a large number of in vitro passages starting with a high- or low-diversity ancestral population. In both cases, we retrieve a huge variability in nucleotide and protein sequences in the one-mutant neighborhood of most abundant sequences, supporting that, at a local level and in a sequential way, viral populations explore a significant fraction of genotype spaces. Analyses of the probability distributions of genotype and phenotype diversity abundance revealed that the order of magnitude of non-neutral mutations was comparable to that of neutral ones. Finally, we show that the functional form of that distribution can be explained with a neutral model, suggesting that diversity generation is strongly shaped by random substitutions, masking the effect of selection at short time scales.

2. Materials and Methods

2.1. Experimental Adaptation of the Qβ Phage to High Temperatures

Plasmid pBRT7Qβ, which contained the bacteriophage Qβ cDNA cloned in the plasmid pBR322 [34], was used to transform *Escherichia coli* DH5-α, which can express viral genes and perform the assembly of the particles, but cannot be infected, since it lacks the F pili. An overnight culture supernatant obtained from a transformed colony was used to infect *E. coli* Hfr (Hayes) on semi-solid agar at a multiplicity of infection (moi) that allowed the generation of well-separated lysis plaques. Under these conditions, there was a high probability that each lysis plaque was the result of the replication of a single virus during a short number of generations and, therefore, could be considered a biological clone. The viral progeny contained in a randomly chosen lysis plate were extracted by collecting the agar occupied by the plaque and incubated in 1 mL of phage buffer (1 g/L gelatin, 0.05 M Tris-HCl, pH 7.5, and 0.01 M $MgCL_2$) with 50 µL of chloroform for 1 h at 28 °C and 850 rpm in a thermoblock. This was then centrifuged for 15 min at 12,000 rpm, and the supernatant was collected. The clonal population thus prepared was subjected to Sanger sequencing and found to have no mutations with respect to the sequence of the cDNA cloned in pBR322. This clone was considered an analog of the wild-type virus and was named the Qβ wild type (Qβ-wt).

The Qβ-wt virus was used to infect an exponential-phase *E. coli* Hfr culture using an moi of 0.1 in a final volume of 1 mL of NB medium (8 g/L Nutrient Broth from Merck and 5 g/L NaCl). After 2 h of incubation at 37 °C, the culture was treated with 1/20 volumes of chloroform, incubated in a thermoblock for 15 min at 37 °C and 600 rpm, and centrifuged for 10 min at 12,000 rpm. Finally, the supernatant was collected and used to infect a new culture of *E. coli* Hfr in the exponential phase, keeping the moi around 0.1. The process was repeated 25 times, each corresponding to a passage or serial transfer. Populations obtained at passages 2 (P2) and 25 (P25) were the ancestors of the evolutionary lines propagated at 43 °C and analyzed in this work. Propagation at 43 °C took place by following the same procedure as at 37 °C, with the only difference being the replication temperature. Adaptation to 43 °C lasted for 60 triplicate passages, giving rise to three populations of $C43_{P2}$(P60) and three $C43_{P25}$(P60) that differed in the initial ancestor. Every ten passages, populations were sequenced through the Sanger method [32]. In this work, we analyzed only one of these triplicates for each population.

2.2. The 3D Structure of the Qβ Replicase

The crystal structure of the Qβ replicase protein is available from the PDB with ID 3MMP [35]. This was the basis for evaluating the structure of our wt sequence (with a 99.83% homology with that in 3MMP), as well as the studied mutants (incorporating from one to three changes in protein sequence). First, the 3D protein structure of the Qβ replicase was obtained through homology modeling based on the actual crystal structure by using Swiss-Model [36]. Different models were constructed for the wild-type (wt) ancestral protein and for variants with mutations A33V, V141A, I319M, S350P, I477T, L509I, L509V, L517F, and G531S.

The obtained models were aligned to the wt protein, and the RMSD was obtained using the *matchmaker* tool in Chimera [37], showing minor differences (in all mutants, below 0.6) with respect to the 3MMP model. The transcription-initiation complexes with the viral protein, the bacterial proteins, and RNA were generated from the β subunits produced through homology modeling, and the transcription factors and RNA molecule were generated from the 3AVT crystallographic structure [38] deposited in the PDB. They corresponded to the initial phase of replication.

2.3. Protein Stability Prediction Methods

The folded stable state of a native protein has a difference in Gibbs energy ΔG with respect to the unfolded conformation. Protein stability prediction (PSP) algorithms quantify the difference in the Gibbs energy of a mutated protein with respect to the native protein:

$$\Delta\Delta G = \Delta G(mutant) - \Delta G(wt) \quad (1)$$

Stabilizing mutations, therefore, yield $\Delta\Delta G > 0$, and *vice versa*.

We used six different algorithms to quantify the change in stability in the 9 mutations listed in the previous section: two variants of I-Mutant 3.0 [39] based on a 3D structure (*Structure*) and sequence (*Seq*), MUpro [40], CUPSAT [41], iStable [42], and FoldX [43].

In order to assess the overall statistical effects of mutations on the protein stability, we measured $\Delta\Delta G$ with FoldX in the following three sets: (i) the top 1000 mutants in $C43_{P2}$(P60) and (ii) $C43_{P25}$(P60), with mutants weighted according to their abundance, and (iii) 1000 mutants with two random residue changes, on average, on the reference wild-type protein. The number of mutations for each mutant was drawn from a Poisson distribution with a parameter of $\lambda = 2$. The latter case served as baseline for comparison with the evolved populations in (i) and (ii). Considering that the data (the distribution of mutant abundance) did not follow a normal distribution, Vargha and Delaney's A test was used to compare the distributions of $\Delta\Delta G$.

2.4. Deep Sequencing Analysis of the C43$_{P2}$(P60) and C43$_{P25}$(P60) Populations

Deep sequencing analysis of line 3 of population C43$_{P2}$(P60) and line 1 of C43$_{P25}$(P60) (see [32] for a description of each line) was carried out at the Genomic Unit of *Parque Científico de Madrid*. Viral RNA was extracted and purified with the QIAamp Viral RNA mini kit (QIAGEN). It was amplified by RT-PCR using SuperScript II Reverse Transcriptase (Invitrogen) and Q5 Hot Star High fidelity (Biolabs) with two pairs of specific nucleotide primers that allowed the generation of amplicons I and II [(P1 forward: 5′GGTGCTTTCG-GTAACATTGAG3′ with P1 reverse: 5′TGAATGAAATACACTCAGCCTCAG3′ to amplify from nucleotide position 2123 to 2516 (amplicon I), and P2 forward 5′AGATTTCTTC-TATGGGTAACGG3′ with P3 reverse 5′CCAGTATTAAATCGGCAGGAC3′ to amplify from nucleotide position 3306 to 3623 (amplicon II)]. PCR products were quantified and tested for quality (TapeStation 4200, Agilent Technologies) prior to the Illumina Ultra Deep Sequencing analysis (MiSeq platform, with the 2 × 250 bp mode and v2 chemistry). After removing the sequence of the primers, amplicon I covered from position 2123 to 2516, which overlapped with the sequence of the replicase gene from nucleotide 2352, and amplicon II covered from position 3306 to 3623, which entirely corresponded to the replicase (Figure 1).

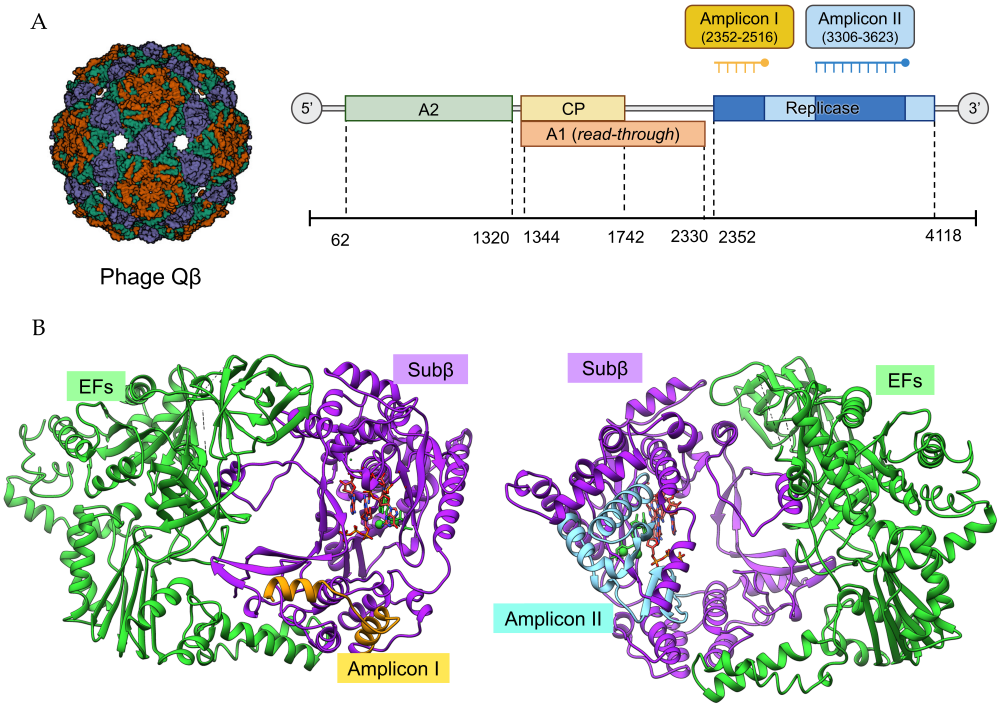

Figure 1. Organization of the Qβ genome and the 3D structure of the virus–cell replicative complex. (**A**) Crystal structure of the capsid of the Qβ phage [44] (left, 1QBE entry in the PDB) and its genome (right). Qβ is an RNA virus with a genome of positive polarity and a capsid composed of about 180 copies of CP protein and a small number of Prot A1 copies, which self-assemble into a 28-nm-diameter icosahedral particle. The genome of Qβ has a length of 4217 nucleotides and contains three open reading frames that encode four proteins: the maturation/lysis protein A2; the coat protein CP; a readthrough of a leaky stop codon in the coat protein, called A1; the β–subunit of an RNA-dependent RNA-polymerase (RdRp), termed the replicase. The two regions overlapping with the proteins that were sequenced (amplicons I and II) are indicated in the picture. (**B**) Two views of the 3D structure of the replicase protein (Subβ unit, in purple) in interaction with the cellular elongation factors (EF, green); see the Materials and Methods for 3D reconstruction methods. We highlight the regions of the replicase corresponding to the folded structure of amplicons I (orange) and II (blue).

2.5. Processing of Deep-Sequencing Data

The Illumina MiSeq Next Generation Sequencer produced high-quality paired-end reads. The following pipeline was applied to the four datasets—amplicons I and II for each population ($C43_{P2}$(P60) and $C43_{P25}$(P60)).

We first performed a quality control of paired-end reads using FastQC v0.11.9 [45] and MultiQC v1.9 [46] to generate a single report for each amplicon in all populations. Cutadapt v3.0 [47] was then used to remove primers, barcode sequences, and low-quality bases (Phred Quality Score < 25) at the 3′ end of forward and reverse reads. Reads that could not be trimmed properly or whose length was smaller than 50 bp were discarded. Since forward and reverse reads overlapped over multiple base pairs, Flash2 v2.2.0 [48] was used to merge paired-end reads in a second step. This step produced high-quality single-end reads [48]. In the third step, we used the BWA v0.7.17 package [49]—specifically, the BWA-MEM aligner algorithm—to map reads against the reference bacteriophage $Q\beta$ genome. Next, in a fourth step, reads from the previous step were sorted and indexed, and unmapped reads were filtered with SAMtools v1.11 [50]. In addition, we used AWK to filter reads that contained either insertions or deletions or whose length was different from that of the reference amplicon. Subsequently, the BAM files were converted into FASTA using SAMtools. Since each of the files corresponding to an experimental population contained a large number of duplicate reads, we used the Collapser tool from the http://hannonlab.cshl.edu/fastx_toolkit/ (accessed on 26 October 2022), FASTX-Toolkit v0.0.14 to collapse duplicated reads into a single read, but while keeping a record of their abundance. All of the steps of the pipeline were implemented in parallel using GNU Parallel [51] and with a customized Bash script. Table 1 shows the number of sequences after data processing and the number of haplotypes (different sequences) obtained after collapsing multiple reads of the same sequence. The scripts used to filter data are available in a Github repository (https://github.com/HSecaira/Quasispecies_TFM/tree/main/Preprocessing (accessed on 26 October 2022)). The two sets of sequences obtained after applying this filtering pipeline to the raw data can be downloaded from the following Github repository: https://github.com/ariadnavillam/QuasispeciesData (accessed on 26 October 2022).

Table 1. Total number of sequences and haplotypes identified through deep sequencing of the two analyzed populations. We show, for each amplicon, the coverage of sequencing (total number of sequences after data processing), the fraction of truncated proteins (those with a non-sense mutation, causing a premature stop codon), the number of different sequences (haplotypes), and the total number of different protein sequences.

Population	Amplicon	Number of Sequences	Truncated Proteins (in %)	Number of Haplotypes	Proteins
$C43_{P2}$(P60)	I	319,825	0.15	7721	4043
	II	364,907	0.5516	20,942	9743
$C43_{P25}$(P60)	I	806,701	0.1631	15,851	8101
	II	692,048	0.4911	34,292	17,094

3. Results

Previous studies have identified multiple non-silent mutations in the $Q\beta$ phage that become highly abundant in the consensus sequence of adapting populations. Specifically, experiments on the adaptation of the $Q\beta$ phage to high temperatures showed that, systematically, the growth rate of the phage, which was initially impaired by the new selective pressure, recovered along adaptive passages. In the 60 passages of the adaptation experiment, however, populations did not attain growth values as high as those in the initial population at 37 °C [32,52,53]. Though specific genomic changes behind adaptation are difficult to disentangle, the fitness of the population, which was measured in those experiments as its growth rate in competition experiments, significantly increases.

Table A1 lists a representative sample of mutations identified through the Sanger methods and some details on the experiments where such mutations emerged. Presumably, they should have played a role in the adaptation of the phage to high temperatures; however, the phenotypic traits affected by such mutations are, so far, unknown. These mutations are used here to evaluate their possible effects on protein stability and to illustrate the difficulties in determining the values of single mutations in heterogeneous genomic and population contexts. These analyses may, at best, offer a partial understanding of the role of isolated non-silent changes. Actually, a clonal analysis of one of such populations adapted to high temperatures showed that mutations never occurred as single mutations in the viral genomes [53]. Thus, their beneficial effects were possibly due to their joint effect on fitness. As a first step towards a more systemic understanding of the evolutionary process, we subsequently studied the genotypic and phenotypic diversity of the phage population along adaptive transients.

3.1. Effects of Single Mutations on Protein Stability

It is customary to explore the effects of non-silent mutations that are fixed in populations of organisms to evaluate their potential adaptive effects. In this section, and as a way of providing an example, we examine the possible effects of several single mutations observed in populations of the Qβ phage adapted to high temperatures (see Appendix A) regarding their effects on molecular structure and protein stability.

Several high-abundance mutations found in the experiments with the phage are represented in Figure 2. None of these were located in the active RdRp site, so these mutations were not expected to affect the interactions with the RNA of the phage, at least at the initial phase of replication. In addition, these mutations were not mutually close in the folded protein, so no direct interactions among them were, in principle, expected. In the following paragraphs, we describe the possible effects of specific mutations found in the adaptation of Qβ to high temperatures.

Residue 33 (Figure 2A) sits in an exposed $\alpha-$helix—though the residue itself is in an internal part—in a hydrophobic environment with amino acids such as Phe and Ile. The substitution Ala $\rightarrow$ Val increases hydrophobicity, probably causing a better fit of its lateral chain. The opposite effect might take place in the substitution V141A (Figure 2B). The affected residue is found in an external turn, and a more exposed chain could enhance its interaction with water molecules. A change with weak *a priori* expected effects is the substitution I319M (Figure 2C) though Met incorporates electric charges due to the electrons of sulfur in the new amino acid.

There are two substitutions that involve hydrogen bonds. The first one is S350P (Figure 2D). The hydroxil (-OH) and amino (-NH3) groups of Ser form two hydrogen bonds with the surrounding amino acids. The residue is located in an exposed turn, a mutation that was also detected in similar structures in mutagenesis studies of the adaptation of enzymes to high temperatures, where Pro enhanced the rigidity of the loop. The second substitution I477T generates a hydrogen bond with a nearby Ser474. Further, Ile (a long-chain amino acid) is substituted by a polar amino acid with a shorter chain.

Other mutations are related to the $\pi-$stacking of $\alpha-$helices, thus improving, in principle, the stability of the folded state. Residue Ile509 sits on an internal $\alpha-$helix surrounded by aromatic residues (Figure 2F). Aromatic rings interact through $\pi-$stacking, causing strong interactions. The observed substitution of Leu by Ile could avoid steric hindrances while maintaining the hydrophobicity of the residue. The original Leu in residue 517 is substituted by Phe, likely enhancing $\pi-$stacking interactions with a nearby Phe (at position 513) (Figure 2G). Though Phe is a hydrophobic residue and position 517 is exposed to the solvent, partial electric charges in the aromatic ring induce the formation of hydrogen bonds with water. A third Phe belonging to ET-Tu factor sits near Phe513, so residues in that region of the replicase might be relevant in the interaction complex. The last abundant mutation entails a substitution of Gly by Ser in residue 531 (Figure 2H). Since this residue

sits in a turn that is fully exposed to the solvent, a hydrophilic amino acid such as Ser interacts with water molecules, potentially improving the stability of the folded state.

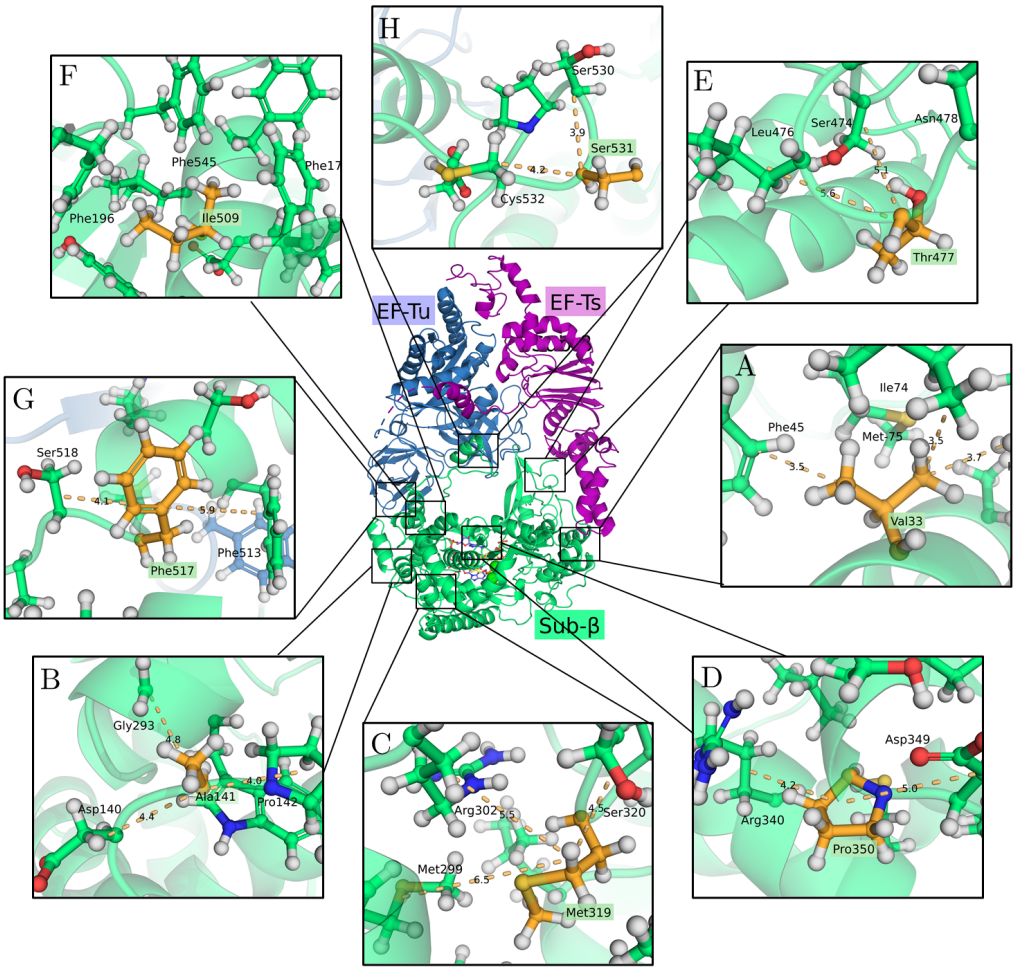

Figure 2. Localization of various non-silent mutations in the structure of the Qβ replicase. Mutations in green were found in the consensus sequences of different populations (see Table A1). We highlighted the closest residues and the distance (in Å) between such residues and the fixed mutation. The new residue is shown in orange and the side chains of the closest amino acids are shown in green. EF: Elongation factors of the cell that belong to the replication complex. Sub-β: β subunit of the viral replicase. As usual, nitrogen atoms appear in blue and oxygen atoms appear in red. (**A**–**H**): See the main text for details on each mutation.

Beyond this qualitative analysis, several PSP methods in the literature allowed a quantitative estimation of the effects of mutations on the folded state of proteins. The reliability of such methods in assessing the effects of single mutations is known to be low, with confidence intervals around $\pm 1 - 1.5 \Delta\Delta G$. Aware of such limitations, we used six of these methods (see Material and Methods) to quantify the predicted changes in free energy caused by the amino acid substitutions observed. All mutations in Appendix A, within one standard deviation when averaging over the six different methods, caused changes within the error interval of the PSP methods used, around or below $\pm 1 \Delta\Delta G$ (see Table A1). In light of the results obtained, we cannot conclude that any of the mutations studied had a clear stabilizing or destabilizing effect on the protein.

Our qualitative and quantitative results with single mutations illustrate the difficulties of assessing the effects of point mutations in isolation, even if only a particular phenotypic trait (protein stability in this case) is taken into account. PSP methods have limited reliability—though, together with independent considerations, they could be indicative of

a sign of specific changes. Expectations thus derived could be improved through molecular dynamics, for example. It is in this way that we obtained independent data supporting the potential stabilizing effect of the mutation I477T [54], for example. It could also be that observable beneficial effects only occur due to the joint action of two or more simultaneously occurring mutations, as experiments measuring population growth rates have shown [53].

3.2. Protein Stability Changes in Ensembles of Mutants

Though stabilizing effects cannot be ascribed to single mutations, previous studies have shown that the analysis of collections of mutants may yield significant trends [55]. Therefore, we explored the overall effects of mutations in our evolved populations. To this end, we calculated changes in folding energy $\Delta\Delta G$ in naturally occurring mutants with respect to the wild-type protein using FoldX. Both amplicons in the two populations were considered and compared to a randomly generated population of mutants with the same number of mutations on average, but they occurred at random positions along the sequence.

Figure 3 summarizes our results. The obtained distributions are notably left-skewed, indicating that mutations, in general, have a destabilizing effect. Since the wild-type protein was likely optimized not only for function, but also for stability throughout evolution, this is a consistent result. However, evolved populations have selected variants that minimize the destabilizing effects of randomly occurring mutations, severely suppressing those with a large negative effect in stability (especially in population $C43_{P2}$(P60)). Indeed, the different dispersions of the distributions in the two evolved populations apparently reflected the different diversities of their ancestral populations. For amplicon I, population $C43_{P25}$(P60) presented values of $\Delta\Delta G$ that were higher than that of the random population, though the difference was not very large. This difference was larger for amplicon II. The differences were larger for population $C43_{P2}$(P60), and, actually, both populations significantly differed when their two amplicons were compared. The values from Vargha and Delaney's A test are reported in Table 2. Summing up, evolved populations increase the stability of the replicase protein overall (or minimize the destabilizing effects of mutations), yielding distributions of changes in folding energy that are significantly narrower than those of random mutants in what can be understood as a signature of positive selection.

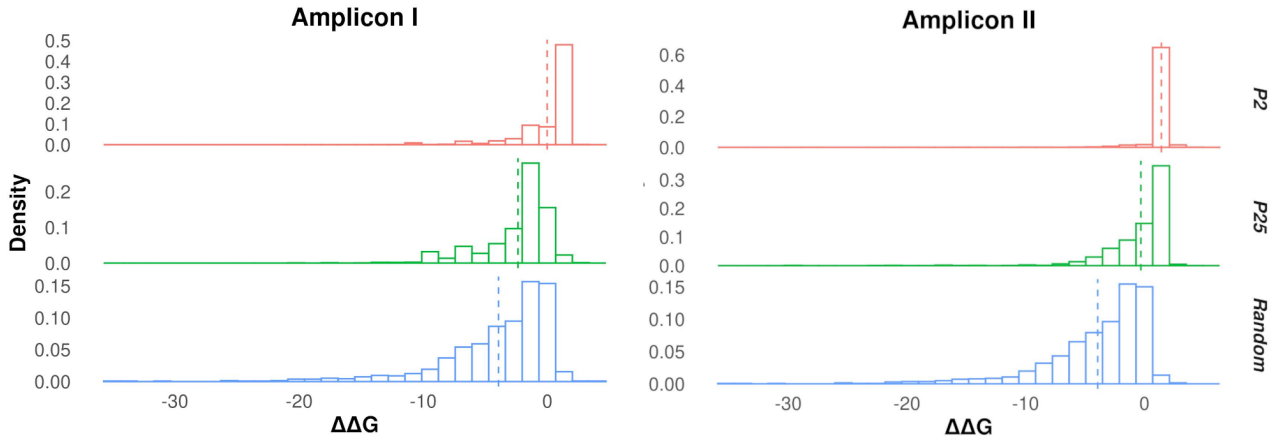

Figure 3. Changes in protein stability for evolved and randomly mutated sequences. The panels show a histogram of changes in folding energy $\Delta\Delta G$ for the two amplicons studied and three different sets of sequences, as indicated. From top to bottom, the distributions correspond to the evolved populations $C43_{P2}$(P60) (labeled P2 in the $y-$axis), $C43_{P25}$(P60) (labeled P25), and 1000 sequences with two (on average) amino acid substitutions at random positions. Mean values are indicated with a dashed line.

Table 2. Vargha and Delaney's A-values for pairwise comparisons of distributions of changes in protein folding energy. An A-value of 0.5 indicates that there is no significant difference.

Amplicon	Populations	A-Value
I	$C43_{P2}$(P60) vs. random	0.87
I	$C43_{P25}$(P60) vs. random	0.61
I	$C43_{P2}$(P60) vs. $C43_{P25}$(P60)	0.82
II	$C43_{P2}$(P60) vs. random	0.98
II	$C43_{P25}$(P60) vs. random	0.83
II	$C43_{P2}$(P60) vs. $C43_{P25}$(P60)	0.93

3.3. Genotypic Diversity of Qβ Populations

A summary of the processed datasets obtained from the two populations of the Qβ phage evolved at 43 °C, with each starting from a different ancestral population (see the Materials and Methods), can be found in Table 1. Current single-molecule sequencing techniques impose a sharp limit on the sequence length that can be faithfully retrieved. Amplicon I had a total length of 328 nucleotides (out of which 121 overlapped with the sequence coding for the replicase protein), while amplicon II reached 274. This yielded replicase (sub)sequences with lengths of 40 and 91 amino acids, respectively. The ensemble of possible different sequences for nucleotides and proteins of equivalent length was attainable neither empirically nor computationally [56]. The figures reached $4^{274} \simeq 10^{165}$ nucleotide sequences or $20^{91} \simeq 10^{118}$ amino acid sequences. These numbers were also indicative of the amount of redundancy induced by the degeneration of the genetic code alone: $10^{165} \gg 10^{118}$; thus, astronomically many different nucleotide sequences must map to the same protein sequence. The fraction of genotype and protein spaces explored by the viral populations and detected through deep-sequencing sampling was, therefore, tiny (see the data in Table 1), on the order of 10^{-161} for genotype sequences and 10^{-87} for protein sequences. Still, the mechanism of natural selection is able to find and fix functional phenotypes in such minute regions, as many previous experiments have shown, and to improve the fitness (growth rate) of Qβ populations [24,30,32,53,57]. These results agree with the expectation that abundant and sufficiently fit phenotypes can be found a few mutations away from any initial sequence.

Figure 4 shows the measured abundances of different haplotypes for ancestral and evolved populations and for the two regions analyzed (amplicons I and II). There are no major differences in the shapes of the four curves shown in each of the two panels, though the total number of haplotypes has almost doubled during evolution (for example, the values for amplicon II of ancestral populations were 13,083 (P2) and 18,188 (P25), which will be compared with the values of evolved populations in Table 1). Ancestral populations appearred to be slightly depleted in terms of highly abundant haplotypes (low-rank values) and rare haplotypes (rightmost part of the curve). Presumably, the population tended to increase the abundance of fitter haplotypes during evolution, with a concomitant exploration of their mutational neighborhood. The shape of the rank distribution, which was compatible with a power-law function in the evolved populations, seemed to be indicative of a hierarchical organization in haplotype abundance. This structure might have just resulted from the replication process of each haplotype together with its diffusion to neighboring haplotypes through random mutations, a hypothesis that could be tested through appropriate dynamical models on genotype networks. In any case, the process of adaptation to high temperatures does not obviously translate into functional modifications of the statistical abundance of haplotypes (but see below).

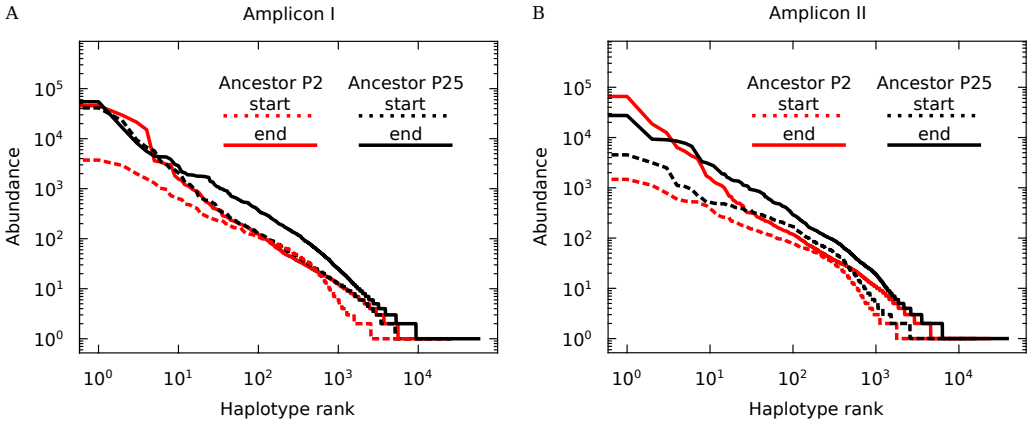

Figure 4. Rank ordering of haplotype abundance for ancestral and evolved populations in the regions corresponding to (**A**) amplicon I and (**B**) amplicon II. See the main text for a discussion.

To complement the above analysis, we calculated the average number of mutations that a haplotype in the population had with respect to the most abundant sequence. To this end, we estimated the frequency of mutations in each amplicon as the ratio between the average number of mutations and the length of the amplicon (the results are in the first column under Nucleotides in Table 3). An estimation of the expected number of mutations that the replicase protein could carry was obtained by multiplying the two independent estimations of mutation frequency (with either replicon) by the total length of the protein. Our results consistently showed that each sequence of nucleotides coding for the replicase in the viral population carried between three and seven mutations (second column under Nucleotides in Table 3), which cause around 2–3 non-silent mutations in the translated protein sequence (columns under Amino Acids in Table 3). These results reveal a high genotypic and phenotypic diversity contained in single populations of the Qβ phage. Protein diversity is further explored in the next section.

Table 3. Mutation frequency in the ensemble of sequences and expected average number of mutations per sequence (nucleotides and amino acids) in the whole protein. Different phenotypes carry multiple amino acid changes when compared to the most abundant protein sequence in the populations.

		Nucleotides		**Amino Acids**	
Population	**Amplicon**	**Frequency**	**Average**	**Frequency**	**Average**
$C43_{P2}$(P60)	I	0.0041	6.916	0.0062	3.68
	II	0.0031	5.235	0.0032	1.89
$C43_{P25}$(P60)	I	0.0022	3.701	0.0038	2.25
	II	0.0022	3.769	0.0033	1.91

3.4. Protein Diversity in Qβ Populations

The distribution of protein abundance in genotype spaces can be analytically calculated according to the limit of independent and uniform site mutation probability in amino acid sequences. This distribution corresponds to the total number of haplotypes that yield the same protein sequence and represents an instance of the distribution of phenotype size (number of genotypes mapping to the same phenotype), as commonly referred to in the literature [8].

Consider an amino acid sequence of length N. Assume that we deal with the $m = 20$ common amino acids listed in Table 4. Let us label these common amino acids, α_i, with an index $i = 1, \ldots, m$, as sorted in the table (while this order is inconsequential). Additionally, let us label amino acids along our sequence, $\alpha(j)$, with a different index $j = 1, \ldots, N$. Note that, at each position j along the sequence, $\{\alpha(j), j = 1, \ldots, N\}$, we have that $\alpha(j) = \alpha_i$ for some i. While the values of α_i are unique, along the sequence, we might find the same amino acid several times. Specifically, let us note as $n_i \leq N$ the number of times that we

find α_i repeated in our sequence. Take as an example the chain GATTACA, with each letter standing for an amino acid, as presented in Table 4. This sequence contains one glycine ($n_{14} = 1$), three alanines ($n_{13} = 3$), two threonines ($n_{16} = 2$), one cysteine ($n_5 = 1$), and none of the other amino acids ($n_i = 0$ for any α_i, $i \notin \{5, 13, 14, 16\}$).

Table 4. List of the 20 most frequent amino acids with their abbreviations. They are ordered according to their versatility (number of different codons for each amino acid) and alphabetically within each versatility value. Three stop codons are not considered, since they correspond to nonsense mutations causing truncated proteins that have been removed from the dataset.

Amino Acid	Abbreviation	One-Letter Abbreviation	i	Versatility, v_i
Methionine	Met	M	1	1
Tryptophan	Trp	W	2	1
Asparagine	Asn	N	3	2
Aspartic acid	Asp	D	4	2
Cysteine	Cys	C	5	2
Glutamine	Gln	Q	6	2
Glutamic acid	Glu	E	7	2
Histidine	His	H	8	2
Lysine	Lys	K	9	2
Phenylanlanine	Phe	F	10	2
Tyrosine	Tyr	Y	11	2
Isoleucine	Ile	I	12	3
Alanine	Ala	A	13	4
Glycine	Gly	G	14	4
Proline	Pro	P	15	4
Threonine	Thr	T	16	4
Valine	Val	V	17	4
Arginige	Arg	R	18	6
Leucine	Leu	L	19	6
Serine	Ser	S	20	6

As remarked above, the space of possible proteins of length N is vast for large values of N. Let us assume that, within this vast space, a small (but also vast) subset of sequences folds into some functional shape. We illustrate this in Figure 5A: The space of all possible combinations of seven amino acids is large (sized $20^7 \sim 10^9$). Within it, we observe two functional sequences: GATTACA and REDRESS. Let us further assume a neutral model—i.e., that all functional sequences have the same fitness, such that biases in how often we find them do not stem from an evolutionary advantage. Finally, let us assume that the $\alpha(j)$ within each functional sequence are uniform and independently drawn from the $m = 20$ amino acids, $\{\alpha_i\}$, and that functional sequences are, in turn, uncorrelated with each other.

In this framework, each functional sequence (each possible sequence, actually) defines a phenotype. While we have assumed fitness neutrality, evolutionary biases might arise from other kinds of constraints. An example is phenotypic size, i.e., the number of different haplotypes that result in a given amino acid sequence. As our data will show, this simple mathematical constraint operates at different scales.

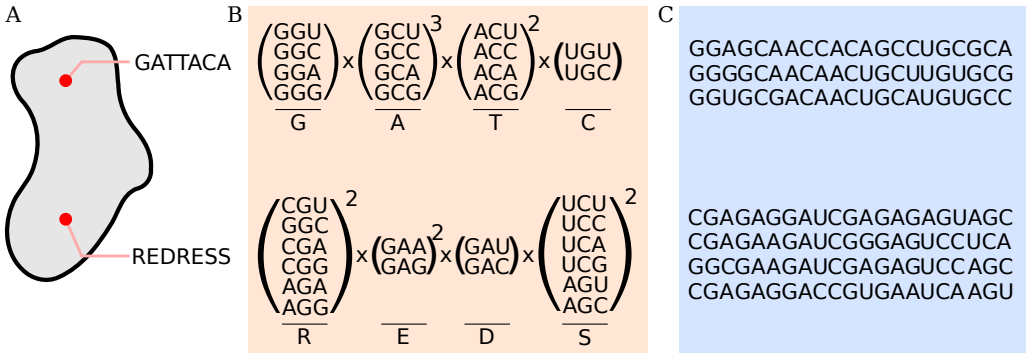

Figure 5. Illustration of phenotypic redundancy in a neutral nucleotide-to-amino-acid-sequence model. (**A**) The space of protein sequences of the same length is vast and contains a variety of functional sequences. (**B**) The number of codons representing amino acid i varies; here, it corresponds to its versatility v_i. Codons coding for glycine (G), alanine (A), threonine (T), cysteine (C), arginine (R), glutamic acid (E), aspartic acid (D), and serine (S) are shown here explicitly as examples. (**C**) Possible nucleotide sequences coding for GATTACA (above) and REDRESS (below).

Following the previous literature [11,16,58], let us consider the *versatility*, v_i, of α_i as the number of different codons that map into that amino acid (summarized in Table 4). The product of the versatilities, $v(j)$, across all positions in a functional sequence gives us the number of different haplotypes that map to that phenotype. This is its phenotypic size. In the case of GATTACA, $\prod_{j=1}^{N} v(j) = 4096$ different genetic sequences result in that same chain of amino acids, while this number increases to $\prod_{j=1}^{N} v(j) = 10{,}368$ for REDRESS (Figure 5B). In general, the phenotypic size S of a protein of a known composition is given by $S \equiv \prod_{j=1}^{N} v(j) = \prod_{i=1}^{m} v_i^{n_i}$.

Taking logarithms, we get:

$$\ln S = \sum_{i=1}^{m} n_i \ln v_i \equiv \sum_{j=1}^{N} \ln v(j). \tag{2}$$

We assumed above that amino acids within a functional sequence are uncorrelated and uniformly distributed. Thus, Equation (2) contains a sum of stochastic variables with a bounded average ($\langle \ln v \rangle \simeq 0.98$) and variance ($\sigma^2_{\langle \ln v \rangle} \simeq 0.27$). Hence, in the large N limit, the frequency of $\ln S$ behaves as a normal distribution with $\mu \equiv N\langle \ln v \rangle \simeq 0.98N$ and $\sigma^2 \equiv N\sigma^2_{\langle \ln v \rangle} \simeq 0.27N$. In other words, we expect the phenotypic size of protein sequences to follow a lognormal distribution with

$$\begin{aligned} \mu_S &= \mu + \frac{\sigma^2}{2} \simeq 1.12N, \\ \sigma_S^2 &= \left(e^{\sigma^2} - 1\right) e^{2\mu+\sigma^2} \simeq \mathrm{e}^{2.51N}. \end{aligned} \tag{3}$$

We plot the theoretical distribution of $\ln S$ in Figure 6A,B (solid, thick blue curves). Alongside this, we plot the distributions of the actual $\ln S$ (as calculated with Equation (2)) for each observed amino acid sequence in all experimental conditions (red histograms for experiments with ancestor P2, black histograms for ancestor P25; Figure 6A for amplicon I, Figure 6B for amplicon II). Leaving aside for a moment the mismatch between both distributions, we note that all observed phenotypes (red and black histograms) lie on the right half of the theoretical abundance curve (solid blue line). This means that only very large phenotypes (i.e., amino acid sequences produced by a vast number of genetic haplotypes) show up in experiments.

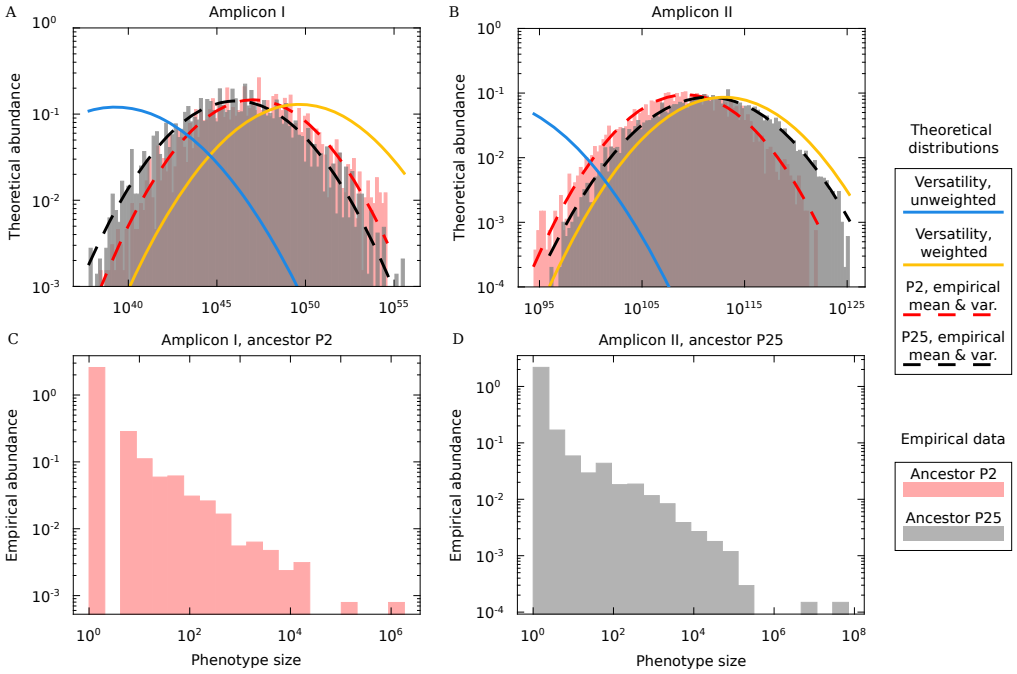

Figure 6. Empirical *versus* theoretical distributions of phenotype sizes. (**A**,**B**) Abundance distribution of protein sequence (phenotype) sizes calculated with various approaches, as shown in the legend and explained in the main text for amplicon I (**A**) and amplicon II (**B**). (**C**,**D**) Abundance distribution of empirical protein sequences calculated as the sum of different empirical haplotypes (nucleotide sequences) that map to the considered protein. All mutations in the nucleotide sequence must, therefore, be silent. These distributions decay from a maximum at size 1 and have fat tails. Two examples are shown: one corresponding to amplicon I that evolved from ancestor P2 (**C**) and one corresponding to amplicon II that evolved from ancestor P25 (**D**).

Let us now tackle the "mismatch" between the theory and experimental data. First, note that the empirical phenotype sizes are well described by a lognormal distribution, nevertheless: the dashed lines in Figure 6A and B show lognormal distributions with the numerical average and variance from the data. This strongly suggests that the abundance of protein sequences in our empirical dataset is well described by a model where amino acids in the sequence are uncorrelated.

Next, let us descend one level deeper—to the empirical genotype sequence—to see how size bias might operate again. Phenotypes are explored by random mutations in the nucleotide chain. Of the large number of genetic haplotypes that map to the same functional sequence (Figure 5B), the exploration of the genotype space can only sample a tiny fraction (Figure 5C). In Figure 6C,D, we show the distributions of genetic haplotypes that were experimentally sampled for each amino acid sequence detected in our experiments. We see that, for most different amino acid sequences, only one genetic haplotype was observed (left side of the empirical abundance distribution).

Actually, we considered that amino acids were equally likely in our calculations above; but, in fact, they were not uniformly sampled. Instead, the process was biased towards those α_i with higher versatility, v_i. The yellow curves in Figure 5A,B show lognormal distributions with average $\mu_\omega \equiv \langle \ln v \rangle_\omega \simeq 1.24N$ and variance $\sigma^2_\omega \equiv \sigma^2_{\langle \ln v \rangle_\omega} \simeq 0.49N$, where the average $\langle \ln v \rangle_\omega$ weights each amino acid proportionally to its versatility. These distributions are closer to the empirical ones.

A discrepancy still remains, as several factors cannot be accounted for in this simple approach. On the one hand, our process is heavily influenced by its initial condition, i.e., an ancestral population with a specific diversity composition, around a highly abundant sequence. This introduces a bias towards the abundances of each amino acid in that original population. On the other hand, while the hypothesis of mutation independence within each

sequence seems to hold, the experimentally observed sequences are highly correlated with each other, since they all stem from a pre-existing parental sequence through replication and mutation. In addition, in the adaptation experiments with the phage, it was the viral quasispecies as a whole that showed adaptation. This violated the assumption of fitness neutrality. Finally, different sequences that are present in the quasispecies might interact with each other, causing cross-interactions and complementation or interference among variants that our simplifying hypotheses cannot capture.

Notwithstanding these and several other overlooked factors, our calculations come quantitatively close to the observed distributions. They are also informative, revealing how phenotype bias operated at two different scales in this experiment. On the one hand, more versatile amino acids were more readily sampled during the exploration of genotype space. This led us to the theoretical distribution with weighted averages (yellow curves in Figure 6A,B). On the other hand, the existence of functional sequences should be a matter of the phenotype space alone. The likelihood that a protein shape is functional cannot depend on the versatility of its constituent amino acids. In other words, assuming fitness neutrality, functional sequences should be sampled uniformly, leading to our first theoretical prediction (blue curves in Figure 6A,B). Tellingly, the parameters (mean and variance) of our empirical distributions lie somewhere in between. However, the fact that the functional form of the distribution remains unchanged supports a dominant role of neutral processes in the evolutionary search.

4. Discussion

The elevated mutation rate that viral genomes experience under replication generates viral populations of high standing diversity, a main adaptive feature of such organisms [24]. Though there might be one to a few sequences of high abundance, the quasispecies is a complex ensemble with many variants of medium and low abundance that are steadily generated through evolution [59]. The maintenance of a variable background of rare mutants is at once the result of selection acting on a heterogeneous population and an evolutionary strategy to carry out a parallel exploration of the many potentially beneficial solutions in new environments.

It is sensible to assume that the increase in abundance of certain variants in a given environment (such as high temperature, in our case) circumstantially supports their adaptive value. This hypothesis can be tested through a variety of methods that hint at possible specific mechanistic effects of certain mutations. However, disentangling in full the cross-effects of a plethora of simultaneous mutations in different genomes remains a challenge for the future.

Studies of the effects of single mutations go beyond the structural analysis performed here and often involve in vitro measures using site-directed mutagenesis. This procedure, however, is both time- and resource-consuming; at least in the case of highly heterogeneous molecular ensembles, such as viruses, it is barely reliable due to the absence of the in vivo molecular environment. Overall, the value of a single mutation depends on its genomic context (the composition of the genome in which it occurs) and on the composition of the accompanying population, with crucial effects in the case of viral quasispecies. For example, neutral mutations that do not change protein sequences can, however, affect the stability of the viral RNA or modify molecular interactions that involve the genomic sequence (as occurs when the genome is packed inside a capsid, or when it replicates using the molecular machinery of the infected cell). Genomic, population, and cellular contexts are impossible to single out for specific viral genomes; this is also a futile effort, since the biological value of genomic changes is meaningless in isolation. On the one hand, and though current techniques permit the determination of the full genomic sequence of individual mutants (through clonal isolation), it is not yet possible to determine the full sequence of a statistically large number of individual variants within a viral population. On the other hand, the effect of a specific mutation cannot be measured in the absence of

the population context due to the many interactions occurring among variants within a quasispecies.

The rapid adaptation of viruses to new environments occurs concomitantly with genomic changes not only due to specific point mutations that rise to fixation in the consensus sequences, but also to collective changes in genome space [60,61]. This is one of the main tenets of this contribution, where we have illustrated the difficulty (and, perhaps, impossibility) of assessing the effects of single mutations, even if they are fixed in the consensus sequence of a viral population. In turn, statistical measures might yield a more reliable signal of adaptation. The search through random mutations is compatible with a neutral process in which mutations would have no effect on fitness while, at the same time, indications of adaptation are retrieved when a phenotypic trait (protein stability) is measured. That is, a population may undergo high diversification during evolution (compatible with neutrality) while, at the same time, it picks out fitter phenotypes (compatible with selection). This reflects the vastness of genotype and phenotype spaces, where the simultaneous occurrence of both facts strongly suggests that multiple haplotypes (and, likely, multiple evolutionary pathways) can promote adaptation.

Our results also support the hypothesis that only very large phenotypes (in our case, those amino acid sequences produced by a vast number of genetic haplotypes) show up in experiments and, arguably, throughout evolutionary history. This observation is in agreement with previous studies addressing the main role of phenotypic bias in evolution, since natural selection seems able to fix only highly abundant phenotypes even if fitter, but less abundant, alternatives exist [62–65].

We are far from fully understanding how genomic changes in complex heterogeneous populations cause adaptive changes in phenotypic traits [8]. Adaptation of heterogeneous molecular populations to new environments is a collective process in which the parallel search of genotype spaces becomes deeply intermingled with the identification of fitter states. The latter states are analogous to emergent properties and, as such, cannot be reduced to the sum of individual contributions: The collective state results from non-trivial interactions among mutants and as a nonlinear addition of minor changes (beneficial, neutral, or even deleterious at the individual level) that turn out to improve fitness at the population level.

Author Contributions: Conceptualization, S.M.; methodology, investigation, formal analysis and validation, A.V., H.S.-M., L.F.S., E.L. and S.M.; software, A.V. and H.S.-M.; resources, E.L. and S.M.; data curation, A.V., H.S.-M. and L.F.S.; writing—original draft preparation, A.V. and S.M.; writing—review and editing, A.V., H.S.-M., L.F.S., E.L. and S.M.; visualization, A.V. and L.F.S.; supervision, E.L. and S.M.; funding acquisition, E.L. and S.M. All authors have read and agreed to the published version of the manuscript.

Funding: The authors are indebted to Luis F. Pacios for his help with the protein stability analyses. The authors acknowledge the financial support from the Spanish State Research Agency, AEI/10.13039/501100011033, through the "Severo Ochoa" Programme for Centers of Excellence in R&D, grant SEV-2017-0712 (A.V., H.S.M., L.F.S., and S.M.), and grants PID2020-113284GB-C21 (L.F.S., S.M.) and PID2020-113284GB-C22 (E.L.). CSIC supported this research through grants JAEINT20-EX-0735 (H.S.M.) and JAESOMdM-2133 (A.V.).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The scripts used to filter the data are available from the following Github repository: https://github.com/HSecaira/Quasispecies_TFM/tree/main/Preprocessing (accessed on 26 October 2022). The two sets of sequences obtained after applying this filtering pipeline to the raw data can be downloaded from the following Github repository: https://github.com/ariadnavillam/QuasispeciesData (accessed on 26 October 2022).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

GP	Genotype-to-phenotype map
PDB	Protein Data Bank
PSP	Protein structure prediction
RdRp	RNA-dependent RNA polymerase
RMSD	Root-mean-square deviation

Appendix A. Mutations in the Qβ Replicase Detected in the Consensus Sequence

In previous experiments of adaptation of the Qβ phage to various conditions, several mutations in the Qβ replicase rose to high abundance, as detected through Sanger sequencing in the consensus sequence. Table A1 lists these mutations, which were used in our study of protein stability described in Section 3.1, and the evolutionary lines in which they were detected. All mutations obtained in the adaptation experiments at 43 °C were reported in [32], with the exception of mutations U3784C and C3879A, which were detected in other experiments of adaptation to 43 °C [24,30,53,57].

Table A1. High-abundance non-synonymous mutations in the replicase of the Qβ phage adapted to 43 °C. Experiments were performed in triplicate, and L1, L2, and L3 indicate that each of the three independent lines simultaneously evolved. The initial Met amino acid is not considered in the enumeration of amino acid positions. The last column shows the average and standard deviation of changes in the folding free energy of the protein with that mutation with respect to the sequence without it, estimated through six different PSP algorithms. See the main text and the Materials and Methods for details.

Nucleotide Mutation	Amino Acid Mutation	Evolutionary Line	ΔΔG
C2452U	A33V	$C43_{P2}$, L1	0.31 ± 0.79
U2776C	V141A	$C43_{P2}$, L1 and L3	-1.24 ± 0.43
U3311G	I319M	$C43_{P2}$, L2; $C43_{P25}$, L1, L2 and L3	-1.07 ± 0.70
U3402C	S350P	$C43_{P2}$, L3	0.14 ± 1.02
U3784C	I477T	Occasional on adaptation to 43 °C	-1.45 ± 0.94
C3879A	L509I	Frequent on adaptation to 43 °C	-0.64 ± 0.63
C3903U	L517F	$C43_{P2}$, L2; $C43_{P25}$, L2 and L3	-1.21 ± 0.77
G3945A	G531S	$C43_{P2}$, L1 and L2; $C43_{P25}$, L1, L2 and L3	-0.80 ± 0.88

References

1. Karlin, S. Some Mathematical Models of Population Genetics. *Am. Math. Mon.* **1972**, *79*, 699–739. [CrossRef]
2. Lanchier, N. Wright–Fisher and Moran models. In *Stochastic Modeling*; Springer International Publishing: Cham, Switzerland, 2017; pp. 203–218. doi: 10.1007/978-3-319-50038-6_12. [CrossRef]
3. Kimura, M. Evolutionary rate at the molecular level. *Nature* **1968**, *217*, 624–626. [CrossRef] [PubMed]
4. Kimura, M. *The Neutral Theory of Molecular Evolution*; Cambridge University Press: Cambridge, UK, 1984.
5. Salisbury, F.B. Natural Selection and the Complexity of the Gene. *Nature* **1969**, *224*, 342–343. [CrossRef] [PubMed]
6. Maynard Smith, J. Natural selection and the concept of a protein space. *Nature* **1970**, *225*, 563–564. [CrossRef]
7. Stadler, P.F.; Stadler, B.M.R. Genotype-Phenotype Maps. *Biol. Theor.* **2006**, *1*, 268–279. [CrossRef]
8. Manrubia, S.; Cuesta, J.A.; Aguirre, J.; Ahnert, S.E.; Altenberg, L.; Cano, A.V.; Catalán, P.; Diaz-Uriarte, R.; Elena, S.F.; García-Martín, J.A.; et al. From genotypes to organisms: State-of-the-art and perspectives of a cornerstone in evolutionary dynamics. *Phys. Life Rev.* **2021**, *38*, 55–106. doi: 10.1016/j.plrev.2021.03.004. [CrossRef]
9. Wagner, A. *The Origins of Evolutionary Innovations*; Oxford University Press: Oxford, UK, 2011.
10. Ahnert, S.E. Structural properties of genotype-phenotype maps. *J. R. Soc. Interface* **2017**, *14*, 20170275. [CrossRef]
11. Garcia-Martin, J.A.; Catalán, P.; Cuesta, J.A.; Manrubia, S. Statistical theory of phenotype abundance distributions: A test through exact enumeration of genotype spaces. *Europhys. Lett.* **2018**, *123*, 28001.
12. Schuster, P.; Fontana, W.; Stadler, P.F.; Hofacker, I.L. From sequences to shapes and back: A case study in RNA secondary structures. *Proc. R. Soc. Lond. B* **1994**, *255*, 279–284.
13. Lipman, D.J.; Wilbur, W.J. Modelling neutral and selective evolution of protein folding. *Proc. R. Soc. Lond. B* **1991**, *245*, 7–11.

14. Jörg, T.; Martin, O.C.; Wagner, A. Neutral network sizes of biological RNA molecules can be computed and are not atypically small. *BMC Bioinform.* **2008**, *9*, 464. [CrossRef] [PubMed]
15. Schaper, S.; Louis, A.A. The arrival of the frequent: How bias in genotype-phenotype maps can steer populations to local optima. *PLoS ONE* **2014**, *9*, e86635.
16. Manrubia, S.; Cuesta, J.A. Distribution of genotype network sizes in sequence-to-structure genotype-phenotype maps. *J. R. Soc. Interface* **2017**, *14*, 20160976. [CrossRef] [PubMed]
17. Greenbury, S.F.; Schaper, S.; Ahnert, S.E.; Louis, A.A. Genetic correlations greatly increase mutational robustness and can both reduce and enhance evolvability. *PLoS Comput. Biol.* **2016**, *12*, e1004773. [CrossRef] [PubMed]
18. Aguirre, J.; Buldú, J.M.; Stich, M.; Manrubia, S.C. Topological structure of the space of phenotypes: The case of RNA neutral networks. *PLoS ONE* **2011**, *6*, e26324. [CrossRef]
19. Yubero, P.; Manrubia, S.; Aguirre, J. The space of genotypes is a network of networks: Implications for evolutionary and extinction dynamics. *Sci. Rep.* **2017**, *7*, 13813. [CrossRef]
20. Grüner, W.; Giegerich, R.; Strothmann, D.; Reidys, C.; Weber, J.; Hofacker, I.L.; Stadler, P.F.; Schuster, P. Analysis of RNA sequence structure maps by exhaustive enumeration II. Structures of neutral networks and shape space covering. *Monatsh. Chem.* **1996**, *127*, 375–389. [CrossRef]
21. Aguilar-Rodríguez, J.; Peel, L.; Stella, M.; Wagner, A.; Payne, J.L. The architecture of an empirical genotype-phenotype map. *Evolution* **2018**, *72*, 1242–1260. [CrossRef]
22. Huynen, M.A.; Stadler, P.F.; Fontana, W. Smoothness within ruggedness: The role of neutrality in adaptation. *Proc. Natl. Acad. Sci. USA* **1996**, *93*, 397–401. [CrossRef]
23. Aguirre, J.; Manrubia, S. Tipping points and early warning signals in the genomic composition of populations induced by environmental changes. *Sci. Rep.* **2015**, *5*, 9664. [CrossRef]
24. Somovilla, P.; Rodríguez-Moreno, A.; Arribas, M.; Manrubia, S.; Lázaro, E. Standing Genetic Diversity and Transmission Bottleneck Size Drive Adaptation in Bacteriophage Qβ. *Int. J. Mol. Sci.* **2022**, *23*, 8876. doi: 10.3390/ijms23168876. [CrossRef] [PubMed]
25. Sanjuán, R.; Nebot, M.R.; Chirico, N.; Mansky, L.M.; Belshaw, R. Viral Mutation Rates. *J. Virol.* **2010**, *84*, 9733–9748. [CrossRef]
26. Harvey, W.; Carabelli, A.; Jackson, B.; Gupta, R.; Thomson, E.; Harrison, E.; Ludden, C.; Reeve, R.; Rambaut, A.; Peacock, S.; et al. SARS-CoV-2 variants, spike mutations and immune escape. *Nat. Rev. Microbiol.* **2021**, *19*, 409–424. doi: 10.1038/s41579-021-00573-0. [CrossRef]
27. Domingo, E.; Sabo, D.; Taniguchi, T.; Weissmann, C. Nucleotide sequence heterogeneity of an RNA phage population. *Cell* **1978**, *13*, 735–744. doi: 10.1016/0092-8674(78)90223-4. [CrossRef]
28. Domingo, E. (Ed.) *Quasispecies: Concept and Implications for Virology*; Springer: Berlin, Germany, 2006.
29. Inomata, T.; Kimura, H.; Hayasaka, H.; Shiozaki, A.; Fujita, Y.; Kashiwagi, A. Quantitative comparison of the RNA bacteriophage Qβ infection cycle in rich and minimal media. *Arch. Virol.* **2012**, *157*, 2163–2169. doi: 10.1007/s00705-012-1419-3. [CrossRef] [PubMed]
30. Kashiwagi, A.; Sugawara, R.; Tsushima, F.S.; Kumagai, T.; Yomo, T.; Simon, A. Contribution of Silent Mutations to Thermal Adaptation of RNA Bacteriophage Qβ. *J. Virol.* **2014**, *88*, 11459–11468. doi: 10.1128/JVI.01127-14. [CrossRef] [PubMed]
31. Lázaro, E.; Arribas, M.; Cabanillas, L.; Román, I.; Acosta, E. Evolutionary adaptation of an RNA bacteriophage to the simultaneous increase in the within-host and extracellular temperatures. *Sci. Rep.* **2018**, *8*, 8080. doi: 10.1038/s41598-018-26443-z. [CrossRef]
32. Somovilla, P.; Manrubia, S.; Lázaro, E. Evolutionary Dynamics in the RNA Bacteriophage Qβ Depends on the Pattern of Change in Selective Pressures. *Pathogens* **2019**, *8*, 80. doi: 10.3390/pathogens8020080. [CrossRef]
33. Arribas, M.; Aguirre, J.; Manrubia, S.; Lázaro, E. Differences in adaptive dynamics determine the success of virus variants that propagate together. *Virus Evol.* **2018**, *4*, vex043. doi: 10.1093/ve/vex043. [CrossRef]
34. Taniguchi, T.; Palmieri, M.; Weissmann, C. QB DNA-containing hybrid plasmids giving rise to QB phage formation in the bacterial host. *Nature* **1978**, *274*, 223–228. doi: 10.1038/274223a0. [CrossRef]
35. Kidmose, R.T.; Vasiliev, N.N.; Chetverin, A.B.; Andersen, G.R.; Knudsen, C.R. Structure of the Qβ replicase, an RNA-dependent RNA polymerase consisting of viral and host proteins. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 10884–10889. doi: 10.1073/pnas.1003015107. [CrossRef] [PubMed]
36. Waterhouse, A.; Bertoni, M.; Bienert, S.; Studer, G.; Tauriello, G.; Gumienny, R.; Heer, F.T.; Beer, T.A.D.; Rempfer, C.; Bordoli, L.; et al. SWISS-MODEL: Homology modelling of protein structures and complexes. *Nucleic Acids Res.* **2018**, *46*, W296–W303. doi: 10.1093/nar/gky427. [CrossRef] [PubMed]
37. Pettersen, E.F.; Goddard, T.D.; Huang, C.C.; Couch, G.S.; Greenblatt, D.M.; Meng, E.C.; Ferrin, T.E. UCSF Chimera - A visualization system for exploratory research and analysis. *J. Comput. Chem.* **2004**, *25*, 1605–1612. doi: 10.1002/jcc.20084. [CrossRef] [PubMed]
38. Takeshita, D.; Tomita, K. Assembly of Qβ viral RNA polymerase with host translational elongation factors EF-Tu and -Ts. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 15733–15738. doi: 10.1073/pnas.1006559107. [CrossRef] [PubMed]
39. Capriotti, E.; Fariselli, P.; Casadio, R. I-Mutant2.0: Predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res.* **2005**, *33*, W306–W310. doi: 10.1093/nar/gki375. [CrossRef]
40. Cheng, J.; Randall, A.; Baldi, P. Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins Struct. Funct. Genet.* **2006**, *62*, 1125–1132. doi: 10.1002/prot.20810. [CrossRef]

41. Parthiban, V.; Gromiha, M.M.; Schomburg, D. CUPSAT: Prediction of protein stability upon point mutations. *Nucleic Acids Res.* **2006**, *34*, W239–W242. doi: 10.1093/nar/gkl190. [CrossRef]
42. Chen, C.W.; Lin, J.; Chu, Y.W. iStable: Off-the-shelf predictor integration for predicting protein stability changes. *BMC Bioinform.* **2013**, *14*, S5. doi: 10.1186/1471-2105-14-S2-S5. [CrossRef]
43. Schymkowitz, J.; Borg, J.; Stricher, F.; Nys, R.; Rousseau, F.; Serrano, L. The FoldX web server: An online force field. *Nucleic Acids Res.* **2005**, *33*, W382–W388. doi: 10.1093/nar/gki387. [CrossRef]
44. Golmohammadi, R.; Fridborg, K.; Bundule, M.; Valegård, K.; Liljas, L. The crystal structure of bacteriophage Qβ at 3.5 Åresolution. *Structure* **1996**, *4*, 543–554. doi: 10.1016/S0969-2126(96)00060-3. [CrossRef]
45. Andrews, S.; Krueger, F.; Segonds-Pichon, A.; Biggins, L.; Krueger, C.; Wingett, S. *FastQC*; Babraham Institute: Cambridge, UK, 2012.
46. Ewels, P.; Magnusson, M.; Lundin, S.; Käller, M. MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **2016**, *32*, 3047–3048. [CrossRef]
47. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* **2011**, *17*, 10–12. [CrossRef]
48. Magoč, T.; Salzberg, S.L. FLASH: Fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **2011**, *27*, 2957–2963. [CrossRef] [PubMed]
49. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* **2013**, arXiv:1303.3997.
50. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R. The sequence alignment/map format and SAMtools. *Bioinformatics* **2009**, *25*, 2078–2079. [CrossRef] [PubMed]
51. Tange, O. Gnu parallel-the command-line power tool. *USENIX Mag.* **2011**, *36*, 42–47.
52. Arribas, M.; Cabanillas, L.; Kubota, K.; Lázaro, E. Impact of increased mutagenesis on adaptation to high temperature in bacteriophage Qβ. *Virology* **2016**, *497*, 163–170. doi: 10.1016/j.virol.2016.07.007. [CrossRef]
53. Arribas, M.; Lázaro, E. Intra-Population Competition during Adaptation to Increased Temperature in an RNA Bacteriophage. *Int. J. Mol. Sci.* **2021**, *22*, 6815. doi: 10.3390/ijms22136815. [CrossRef]
54. Villanueva Marijuán, A. Análisis de las Mutaciones en la Replicasa del Virus Qβ, 2021. TFG Thesis. Available online: https://oa.upm.es/69689/ (accessed on 26 October 2022).
55. Potapov, V.; Cohen, M.; Schreiber, G. Assessing computational methods for predicting protein stability upon mutation: Good on average but not in the details. *Protein Eng. Des. Sel.* **2009**, *22*, 553–560. doi: 10.1093/protein/gzp030. [CrossRef]
56. Louis, A.A. Contingency, convergence and hyper-astronomical numbers in biological evolution. *Stud. Hist. Philos. Sci. C* **2016**, *58*, 107–116. [CrossRef]
57. Arribas, M.; Kubota, K.; Cabanillas, L.; Lázaro, E. Adaptation to Fluctuating Temperatures in an RNA Virus Is Driven by the Most Stringent Selective Pressure. *PLoS ONE* **2014**, *9*, e100940. doi: 10.1371/journal.pone.0100940. [CrossRef] [PubMed]
58. Cuesta, J.A.; Manrubia, S. Enumerating secondary structures and structural moieties for circular RNAs. *J. Theor. Biol.* **2017**, *419*, 375–382. [CrossRef] [PubMed]
59. Domingo, E.; Soria, M.E.; Gallego, I.; de Ávila, A.I.; García-Crespo, C.; Martínez-González, B.; Gómez, J.; Briones, C.; Gregori, J.; Quer, J.; et al. A new implication of quasispecies dynamics: Broad virus diversification in absence of external perturbations. *Infect. Genet. Evol.* **2020**, *82*, 104278. doi: 10.1016/j.meegid.2020.104278. [CrossRef]
60. Perales, C.; Henry, M.; Domingo, E.; Wain-Hobson, S.; Vartanian, J.P. Lethal Mutagenesis of Foot-and-Mouth Disease Virus Involves Shifts in Sequence Space. *J. Virol.* **2011**, *85*, 12227–12240. doi: 10.1128/JVI.00716-11. [CrossRef] [PubMed]
61. Agudo, R.; Ferrer-Orta, C.; Arias, A.; de la Higuera, I.; Perales, C.; Pérez-Luque, R.; Verdaguer, N.; Domingo, E. A Multi-Step Process of Viral Adaptation to a Mutagenic Nucleoside Analogue by Modulation of Transition Types Leads to Extinction-Escape. *PLoS Pathog.* **2010**, *6*, e1001072. doi: 10.1371/journal.ppat.1001072. [CrossRef] [PubMed]
62. Cowperthwaite, M.C.; Economo, E.P.; Harcombe, W.R.; Miller, E.L.; Meyers, L.A. The Ascent of the Abundant: How Mutational Networks Constrain Evolution. *PLoS Comput. Biol.* **2008**, *4*, e1000110. [CrossRef] [PubMed]
63. Dingle, K.; Schaper, S.; Louis, A.A. The structure of the genotype-phenotype map strongly constrains the evolution of non-coding RNA. *Interface Focus* **2015**, *5*, 20150053. [CrossRef]
64. Catalán, P.; Manrubia, S.; Cuesta, J.A. Populations of genetic circuits are unable to find the fittest solution in a multilevel genotype-phenotype map. *J. R. Soc. Interface* **2020**, *17*, 20190843. [CrossRef]
65. Dingle, K.; Ghaddar, F.; Šulc, P.; Louis, A.A. Phenotype bias determines how natural RNA structures occupy the morphospace of all possible shapes. *Mol. Biol. Evol.* **2022**, *39*, msab280. [CrossRef]