Biochemistry

Article

Cite This: Biochemistry 2019, 58, 1527−1538

pubs.acs.org/biochemistry

Experimental Evaluation of Coevolution in a Self-Assembling Particle

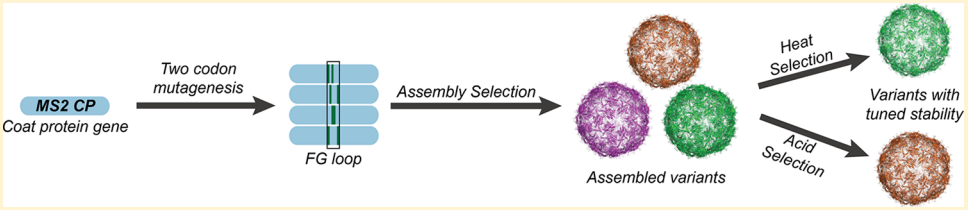
Emily C. Hartman,[†] Marco J. Lobba,[†] Andrew H. Favor,[†] Stephanie A. Robinson,[†] Matthew B. Francis,*[,†,‡] and Danielle Tullman-Ercek*[,§]

[†]Department of Chemistry, University of California, Berkeley, California 94720-1460, United States

[‡]Materials Sciences Division, Lawrence Berkeley National Laboratories, Berkeley, California 94720-1460, United States

[§]Department of Chemical and Biological Engineering, Northwestern University, 2145 Sheridan Road, Technological Institute E136, Evanston, Illinois 60208-3120, United States

Supporting Information



ABSTRACT: Protein evolution occurs via restricted evolutionary paths that are influenced by both previous and subsequent mutations. This effect, termed epistasis, is critical in population genetics, drug resistance, and immune escape; however, the effect of epistasis on the level of protein fitness is less well characterized. We generated and characterized a 6615-member library of all two-amino acid combinations in a highly mutable loop of a virus-like particle. This particle is a model of protein self-assembly and a promising vehicle for drug delivery and imaging. In addition to characterizing the effect of all double mutants on assembly, thermostability, and acid stability, we observed many instances of epistasis, in which combinations of mutations are either more deleterious or more beneficial than expected. These results were used to generate rules governing the effects of multiple mutations on the self-assembly of the virus-like particle.

Protein evolution occurs through complex pathways, often involving nonintuitive leaps between functional variants.[1−3] These paths include local minima and maxima, in which the effect of a given mutation depends entirely on the previous and subsequent mutation.[1] This effect, known as epistasis, has been well-studied in population genetics[4−6] and is known to play a central role in drug resistance[7,8] and immune escape.[4,6,9] However, studies that quantify the combinatorial effect of multiple mutations on protein fitness remain relatively rare.

Much effort has focused on characterizing the fitness effect of single mutations on a given protein, producing one-dimensional protein fitness landscapes.[10−15] While such landscapes are highly useful for describing the effects of one-amino acid mutations on a protein, these efforts do not capture the multidimensional shape (or ruggedness) of evolutionary landscapes. To date, epistasis has been measured for GFP,[16] RNA-binding proteins,[17] and several enzymes.[2,18,19] These studies find that instances of negative epistasis, in which a secondary mutation is more deleterious than anticipated, are more common than positive epistasis, though both are detectable and play a role in shaping protein fitness landscapes.[16,17,20−22]

Complex, multimeric protein scaffolds such as viral capsids, metabolosomes, and other molecular machines are poised to have a significant impact on biotechnology in the coming decades.[23,24] To maximize this potential, it is important to understand how non-native functions can be hindered by unanticipated epistatic effects. To date, our understanding of the design rules governing the self-assembly of these proteins remains limited, complicating the use, predictability, and yield of these particles in non-native contexts.[25,26] In particular, we expect the effects of epistasis to be especially significant in large assemblies with quaternary structure due to many inter- and intramonomer interactions.[27,28]

Here, we characterize the complete pairwise epistasis of a highly mutable loop occurring in a virus-like particle (VLP) (Figure 1). VLPs are closed-shell protein containers derived from noninfectious viral shell proteins. MS2 bacteriophage is used as a model VLP, as it is well-studied for use in drug delivery,[29−31] disease imaging,[32−34] vaccine development,[35,36] and biomaterials.[37,38] To do this, we harness a technique developed in our laboratories, called SyMAPS[15] (systematic mutation and assembled particle selection), to evaluate how

Special Issue: The Chemistry of Synthetic Biology

Received: September 5, 2018
Revised: November 2, 2018
Published: November 12, 2018

DOI: 10.1021/acs.biochem.8b00948
Biochemistry 2019, 58, 1527−1538

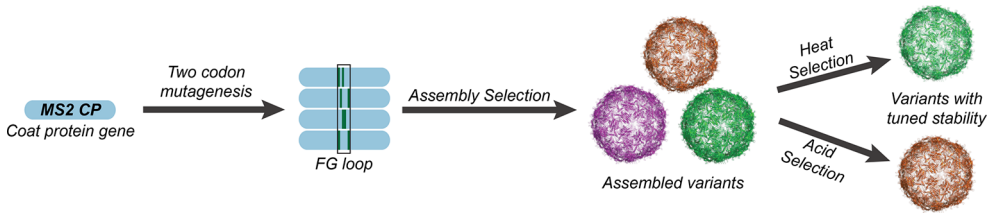


Figure 1. FG loop mutagenesis and selection strategy. Two-codon mutagenesis targeted at the FG loop generated a library of 6615 variants, each with exactly two NNK substitutions in this region. These variants were subjected to an assembly selection, followed by targeted functional selections to identify heat-stable and acid-sensitive variants.

epistasis shapes the assembly, thermostability, and acid stability of the MS2 coat protein (CP). On the basis of our selection criteria, we find many instances of both positive and negative epistasis for the loop region studied, governed largely by charge and steric bulk. Our studies reveal two residues, distant by sequence but spatially adjacent, that show strong pairwise epistasis. This allows us to describe unexpected design rules governing the mutability of this loop. Moreover, this work establishes a useful experimental protocol that can be used to understand how epistatic effects can be leveraged to obtain new particles with desired physical or chemical features.

MATERIALS AND EXPERIMENTAL DETAILS

Strains. MegaX DH10B *Escherichia coli* electrocompetent cells (ThermoFisher Scientific, catalog no. C640003) were used for all library experiments, and DH10B chemically competent cells produced in house were used for expression of individual variants of interest. Overnight cultures from a single colony were grown for 16−20 h at 37 °C while being shaken at 200 rpm in LB-Miller medium (Fisher Scientific, catalog no. BP1426-2) with 32 mg/L chloramphenicol. Expressions were subcultured in a 1:100 ratio into 2×YT medium (Teknova, catalog no. Y0210) with 32 mg/L chloramphenicol and allowed to express overnight at 37 °C while being shaken at 200 rpm.

Library Generation. To generate libraries with two-amino acid mutations in the FG loop, we modified a library generation strategy developed by the Bolon lab, known as EMPIRIC cloning.[39] EMPIRIC cloning uses a plasmid with a self-encoded removable fragment (SERF) surrounded by inverted BsaI restriction sites. With this setup, BsaI digestion simultaneously removes both SERF and BsaI sites. These plasmids are termed entry vectors, and the SERF in this study encodes constitutively expressed GFP to permit green/white screening. We used a previously described entry vector that replaced a 26-codon segment flanking the FG loop in the MS2 CP with the SERF.[15] Single-stranded DNA primers with all 15 combinations of degenerate codons were purchased, enabling overlap extension polymerase chain reaction (PCR) to generate double-stranded DNA with all possible pairwise combinations in this six-residue region. These primers were resuspended in water, pooled, and diluted to a final concentration of 50 ng/μL. The reverse strand was filled in by overlap extension PCR with a corresponding forward primer. The amplified, double-stranded DNA was purified using a PCR Clean-up Kit (Promega, catalog no. A9282). The purified DNA was diluted to 1−5 ng/μL and cloned into the described entry vector using established Golden gate cloning techniques.[40] The ligated plasmids were desalted on membranes (Millipore Sigma, catalog no. VSWP02500) for 20 min and then transformed into MegaX DH10B *E. coli* electrocompetent cells (ThermoFisher Scientific, catalog no. C640003). Following electroporation and recovery, cells were plated onto two large LB-A plates (VWR, catalog no. 82050-600) with 32 μg/mL chloramphenicol and allowed to grow at 37 °C overnight. The colony number varied, but every transformation yielded a number of colonies that was at least 50 times greater than the library size. This protocol was repeated in full for three total biological replicates that are fully independent from library generation through selection.

Size Selection. Colonies were scraped from plates into LB-M and allowed to grow for 2 h. Each library was then subcultured at a 1:100 ratio into 1 L of 2×YT (Teknova, catalog no. Y0210) and allowed to grow to an OD_{600} of 0.6, when they were induced with 0.1% arabinose. Libraries of variants were expressed overnight at 37 °C. Cultures were then harvested, resuspended in 10 mM phosphate buffer (pH 7.2) with 2 mM sodium azide, and sonicated. Libraries were subjected to two rounds of 50% (w/v) ammonium sulfate precipitation, followed by fast protein liquid chromatography (FPLC) size exclusion chromatography purification to select for well-formed VLPs.

FPLC SEC (assembly selection). MS2 CP libraries or individual variants were purified on an Akta Pure 25 L FPLC system with a HiPrep Sephacryl S-500 HR column (GE Healthcare Life Sciences, catalog no. 28935607) size exclusion chromatography (SEC) column via isocratic flow with 10 mM phosphate buffer (pH 7.2), 200 mM sodium chloride, and 2 mM sodium azide. Fractions containing MS2 coat protein were collected for further analysis.

Heat Selection. Following assembly selection (FPLC purification), libraries were incubated at 50 °C for 10 min. The buffer from FPLC purification [10 mM phosphate buffer (pH 7.2), 200 mM sodium chloride, and 2 mM sodium azide] was used for these studies. Precipitated VLPs were pelleted by centrifugation, and well-formed VLPs were isolated by semipreparative high-performance liquid chromatography (HPLC) SEC. Fractions containing VLP were combined and subjected to RNA extraction, barcoding, and high-throughput sequencing.

HPLC SEC. MS2 CP variants were analyzed on an Agilent 1290 Infinity HPLC system with an Agilent Bio SEC-5 column (5 μm, 2000 Å, 7.8 mm × 300 mm) with an isocratic flow of 10 mM phosphate buffer (pH 7.2), 200 mM sodium chloride, and 2 mM sodium azide. Fractions were collected at the characteristic elution time for wild-type MS2 (11.2 min) and subjected to RNA extraction and high-throughput sequencing sample preparation.

Acid Selection. Libraries were incubated at pH 5 or 7 for 1 h at 37 °C, prepared with citric acid and sodium phosphate according to the Sigma-Aldrich Buffer Reference Center. Precipitated VLPs were pelleted by centrifugation, and intact

DOI: 10.1021/acs.biochem.8b00948
Biochemistry 2019, 58, 1527−1538

VLPs were concentrated using a 100 kDa molecular weight spin cutoff filter (Millipore Sigma, catalog no. UFC510024). The supernatant of the MWCO filter was subjected to RNA extraction, barcoding, and high-throughput sequencing as described above.

Sample Preparation for High-Throughput Sequencing. Plasmid DNA was extracted prior to expression using a Zyppy Plasmid Miniprep Kit (Zymo, catalog no. D4036). RNA was extracted from the MS2 CP library following assembly selections using previously published protocols.[41] Briefly, TRIzol (Thermo Fisher, catalog no. 15596026) was used to homogenize samples, followed by chloroform addition. The sample was separated by centrifugation into aqueous, interphase, and organic layers. The aqueous layer, which contained RNA, was isolated, and the RNA was then precipitated with isopropanol and washed with 70% ethanol. RNA was then briefly dried and resuspended in RNase free water. cDNA was then synthesized using the Superscript III first-strand cDNA synthesis kit from Life (catalog no. 18080051, polyT primer). cDNA and plasmids were both amplified with two rounds of PCR to add barcodes (10 cycles) and the Illumina sequencing handles (8 cycles), respectively, following Illumina 16S Metagenomic Sequencing Library Preparation recommendations (Data Set S1). Libraries were combined and analyzed by 150 PE MiSeq in collaboration with the University of California Davis Sequencing Facilities. Reads in excess of 18 million passed filter, and an overall Q30 > 85%.

Individual Variant Cloning. Individual variants were cloned using a variation on the method described above. Briefly, overlap extension PCR (Data Set S1) yielded a double-stranded fragment that spanned the length of the missing 26-codon region in the entry vector. Each fragment was cloned into the entry vector using standard Golden gate cloning techniques.[40] Variants bearing the CP[N87C] mutation were cloned into a similar entry vector bearing the desired mutation at position 87, which was installed via site-directed mutagenesis.[42] Cloned plasmids were transformed into chemically competent DH10B cells. Individual colonies were sequenced via Sanger sequencing prior to expression.

Individual Variant Expression. Selected mutants were individually expressed in 5 or 50 mL cultures of 2×YT. These expressions were pelleted, resuspended in 10 mM phosphate buffer (pH 7.2) and 2 mM sodium azide, lysed by sonication, precipitated twice with 50% (w/v) ammonium sulfate, and evaluated by a native gel for VLP formation and acid sensitivity.

Individual Acid Screens. Following ammonium sulfate precipitation and resuspension, variants were diluted at a 1:10 ratio into neutral or acidic buffers ranging from pH 3.9 to 7.4. These buffers contained various concentrations of citric acid and phosphate, prepared according to the Sigma Buffer Reference. Variants were centrifuged at 13000*g* for 2 min, and then equal volumes were loaded onto a native gel. Densitometry with ImageJ was used to determine VLP formation and sensitivity to acidic conditions.

Native Gel. VLPs were analyzed in a 0.8% agarose gel in 0.5× TBE buffer (45 mM Tris-borate and 1 mM EDTA) with 2× SYBR Safe DNA Gel Stain (ThermoFisher Scientific, catalog no. S33102) for 120 min at 40 V. Agarose gels were imaged on a Bio-Rad GelDoc EZ Imager. Densitometry under each condition compared to pH 7.4 was performed using ImageJ.

Sypro Orange Melting Curves. Individual variants were purified by HPLC SEC and then diluted to a final A_{280} of 1 in 10 mM phosphate buffer (pH 7.2), 200 mM sodium chloride, and 2 mM sodium azide. Purified variants were filtered, and a final concentration of 20× Sypro Orange was added. The melting temperature was determined using a qPCR protocol that measured fluorescence every 0.5 °C between 30 and 80 °C using an Applied Biosystems QuantStudio 3. The derivative was taken of the resulting fluorescence curves, and the minimum value was determined to be the melting temperature.

Cysteine Modification. Variants of interest were purified by HPLC SEC and diluted to 5 μM in 20 mM phosphate buffer (pH 7.2). A solution of AlexaFluor-488 maleimide (ThermoFisher Scientific, catalog no. A10254) in dimethylformamide was added to a final concentration of 20× (relative to capsid monomer) and allowed to react for 1 h. Variants were spin concentrated three times with a centrifugal filter with a 100 kDa molecular weight cutoff (Millipore Sigma, catalog no. UFC510024) and then evaluated by HPLC SEC and ESI TOF.

Mass Spectrometry. Modified and unmodified proteins were analyzed with an Agilent 1200 series liquid chromatograph (Agilent Technologies, Santa Clara, CA) connected in line with an Agilent 6224 time-of flight (TOF) LC/MS system with a Turbospray ion source.

High-Throughput Sequencing Data Analysis. Data were trimmed and processed as previously described[15] with minor variations. Briefly, data were trimmed with Trimmomatic[43] with a four-unit sliding quality window of 20 and a minimum length of 30. Reads were merged using FLASH (fast length adjustment of short reads)[44] with a maximum overlap of 160 bp. Reads were then aligned with the wild-type MS2 CP reference gene with Burrows-Wheeler Aligner (BWA-MEM).[45] Reads were then sorted and indexed with Samtools.[46] The Picard function CleanSam was used to filter unmapped reads, and reads longer or shorter than the expected length of the barcoded DNA were removed.

AFL Calculations. Cleaned and filtered high-throughput sequencing reads were analyzed using Python programs written in house. Briefly, the mutated region of the MS2 CP was isolated, and the number of mutations per read was calculated. Reads with zero mutations (wild-type reads) or more than two mutations were both removed. In reads with two mutations, the two non-wild-type codons were identified and counted. In reads with one mutation, the mutated codon was tallied in combination with every wild-type codon. Codons were then translated into amino acids, removing codons that do not end in G or T.

These calculations were repeated for all experiments to generate abundances before and after each selective pressure. Relative percent abundances were calculated as previously described.[15] Briefly, the grand sum, or the sum of all counts at every combination at every position, was calculated. We next divided each matrix by its grand sum, generating a matrix of percent abundances. These calculations were repeated for each biological replicate of VLP, plasmid, heat-selected, or acid-selected libraries, generating 12 different percent abundance matrices. We calculated relative percent abundances by dividing the percent abundance for the selected library compared by the percent abundance for the plasmid library for each replicate.

We calculated the mean across the three replicates. All Nan (null) values, which indicate variants that were not identified in

DOI: 10.1021/acs.biochem.8b00948
Biochemistry 2019, 58, 1527−1538

the plasmid library, were ignored. Scores of zero, which indicate variants that were sequenced in the unselected library but absent in the VLP library, were replaced with an arbitrary score of 0.0001. We calculated the $\log_{10}$ of the relative percent abundance array to calculate the final array for each replicate. Finally, we calculated the average AFS value for each amino acid combination by finding the mean value for every combination, which is displayed in Figure S1. In addition, all AFS values for assembly and heat selections can be found in Data Set S2.

Shannon Entropy Calculations. The Shannon entropy is defined as

$$\text{Shannon entropy} = -\sum P \log(P) \tag{1}$$

where P refers to a given probability. We first calculate the probability of a given combination of two amino acids occurring within a single residue. Any zero values are replaced with 0.00001. We then calculated the Shannon entropy at all 15 combinations as follows.

Shannon entropy values were averaged across three biological replicates for each library. The difference between the unselected plasmid library and the VLP library, or the unselected plasmid library compared to the heat-selected VLP library, was used to evaluate how mutability affects thermostability.

Predicted Two-Dimensional (2D) AFL: Convolutional Neural Network. All neural network model development was conducted in Python using the Tensorflow library.[47] The neural network design was composed of two sequential convolutional layers, each followed by a pooling operation, that ultimately fed into two fully connected layers, which output a scalar fitness score prediction. Data were fed in as an array consisting of 12 physical properties listed for each amino acid position along the MS2 backbone.[48−52] Mutations were modeled by swapping the physical properties of one amino acid for another at the relevant site in the MS2 CP. Mutants with fitness scores less than or equal to −4 were removed, and missing data points were excluded entirely.

The property columns in our input matrices are separately fed into the function, processed individually, and then combined. The first convolutional layer takes small fragments of the input vector (kernel size = 5), corresponding to five-residue sequences of amino acids in the backbone of the MS2 CP. This length was chosen to represent small units of sequence that can exhibit characteristic patterns in their physical identities without overburdening the training model. After each pass through the filters, a "pooling" operation reduces the size of the data passed along by taking the maximum value of each 2-unit long subdivision of the filter outputs and consolidating them to feed into the next layer. The neural net was trained on the full one-dimensional (1D) AFL using the mean squared error as our optimization factor, with the maximum number of iterations set to 50000.

Epistasis Calculations. Epistasis, E, was calculated as described elsewhere.[16] Briefly, we calculated the difference between nonadditive effects of one-amino acid mutations on a log scale. We first calculated the difference between the AFS value of a mutation (AFS_i) and the average AFS_{WT}. $\text{AFS}_{\text{predicted}}$ was calculated by adding the Δ_i values for each mutation to the average AFS_{WT}. This predicted AFS values for all two-amino acid variants are plotted in Figure S7B. The predicted 2D AFS value was then subtracted from the measured AFS value to generate E, a measure of epistasis.

$$\Delta_i = \text{AFS}_i - \text{AFS}_{\text{WT}} \tag{2}$$

$$\text{AFS}_{\text{predicted}} = \left(\sum_i \Delta_i\right) + \text{AFS}_{\text{WT}} \tag{3}$$

$$E = \text{AFS}_{\text{measured}} - \text{AFS}_{\text{predicted}} \tag{4}$$

Variants where the sign (+ or −) between the predicted 2D AFL and the measured 2D AFL was inverted were separated for further analysis. E values for these variants are plotted in Figure 5A. Epistasis scores are available in Data Set S3.

Phylogeny Calculations. Bacteriophage coat proteins related to the MS2 CP (Protein Data Bank entry: 2MS2) were identified and aligned with UniProt Align, and a phylogenetic tree was calculated using Interactive Tree of Life (iTOL).[53] Consensus sequences were generated using Berkeley WebLogo.[54]

RESULTS AND DISCUSSION

We recently described the 1D AFL of the MS2 CP.[15] In this fitness landscape, we evaluated the mutability at each position across the MS2 CP. While the MS2 CP has been extensively used as an epitope display platform, mutations and peptide insertions are typically performed on an exterior-facing loop between residues 14 and 19.[55,56] In our previous study, we found that a six-amino acid stretch in a flexible loop connecting two β-sheets, termed the FG loop, was highly mutable (Figure 2A), meaning that many amino acid substitutions assembled into well-formed VLPs. This FG loop undergoes a critical conformational shift during VLP assembly,[57−59] which results in two distinct structures near the pore, termed the A/C (Figure 2B) and B (Figure 2C) forms. Because a conformational change in this loop is important for VLP assembly,[57−59] we were surprised by the mutability of this region.

This loop was used as a model system to study how two-amino acid mutations affect protein fitness, thereby characterizing a second dimension of the MS2 CP protein fitness landscape. To evaluate the fitness of all variants in this library, we used SyMAPS, a technique that generates a quantitative score of assembly competency across a targeted library of VLP variants. When expressed in *E. coli*, well-formed particles will encapsulate available negative charge within the MS2 CP during assembly.[60] SyMAPS takes advantage of this property, using intrinsic nucleic acid encapsulation as a convenient genotype-to-phenotype link. If a given MS2 CP mutation assembles, then variant mRNA is encapsulated within the VLP and copurifies with it.[15] If assembly is not permitted with a given mutation, then cellular nucleic acids are not recovered.

We generated a 6615-member library containing all possible two-amino acid combinations in the FG loop (T71−E76). This library contained all single and double amino acid mutations from the native MS2 CP sequence. We subjected the library to a selection based on VLP assembly and performed high-throughput sequencing before and after the selection. The library was generated and analyzed in three independent replicates. The percent abundance of each variant before and after the selection was converted into an apparent fitness score (AFS). Library members with positive AFS values correspond to mutations that permit assembly. Conversely, negative AFS values indicate disfavored VLP formation, which

DOI: 10.1021/acs.biochem.8b00948
Biochemistry 2019, 58, 1527−1538

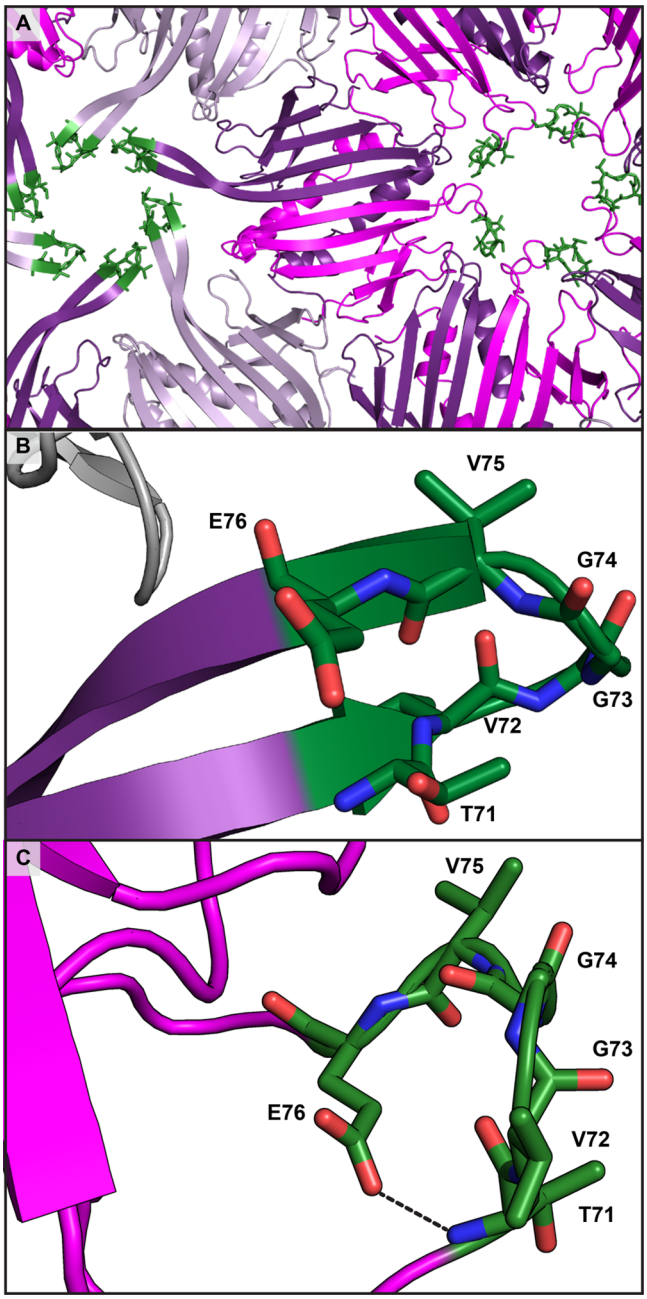


Figure 2. Changes in the FG loop of the MS2 CP. (A) The view from the interior of the capsid shows how the FG loop (green) differs between the quasi-6-fold and 5-fold axes. A-type monomers are colored dark purple, B monomers magenta, and C monomers light purple. Close-up perspectives of the (B) quasi-6-fold and (C) 5-fold axes highlight the structural changes between the A/C and B form monomers, respectively. A key hydrogen bond between E76 and the backbone of T71 is indicated with a dashed line in panel C.

could be due to poor expression, inefficient or no assembly, or instability to protein purification. These scores are presented together in a combined 2D AFL, which indicates the effects of all one- and two-amino acid mutations in this loop on VLP assembly (Figure S1).

Of the 6615 possible library members, >92% of all variants were identified in at least one replicate and 87% were identified in all three replicates. Approximately 5% of the variants were sequenced in the plasmid library but were absent in the VLP library; these variants are colored dark red on the 2D AFL. Nonsense and silent mutations were used as internal negative and positive controls, respectively. In this study, 483 nonsense mutations were measured, and all had an AFS value of −0.19 or lower, correctly identifying each variant as nonassembling. In contrast, 15 silent mutations, or one per combination, were measured, and all of these AFS values were 0.20 or higher. This indicates that these wild-type VLPs are correctly identified as being well-assembled. As expected, these two populations separated into two nonoverlapping groups (Figure 3A), affirming the quality of these data.

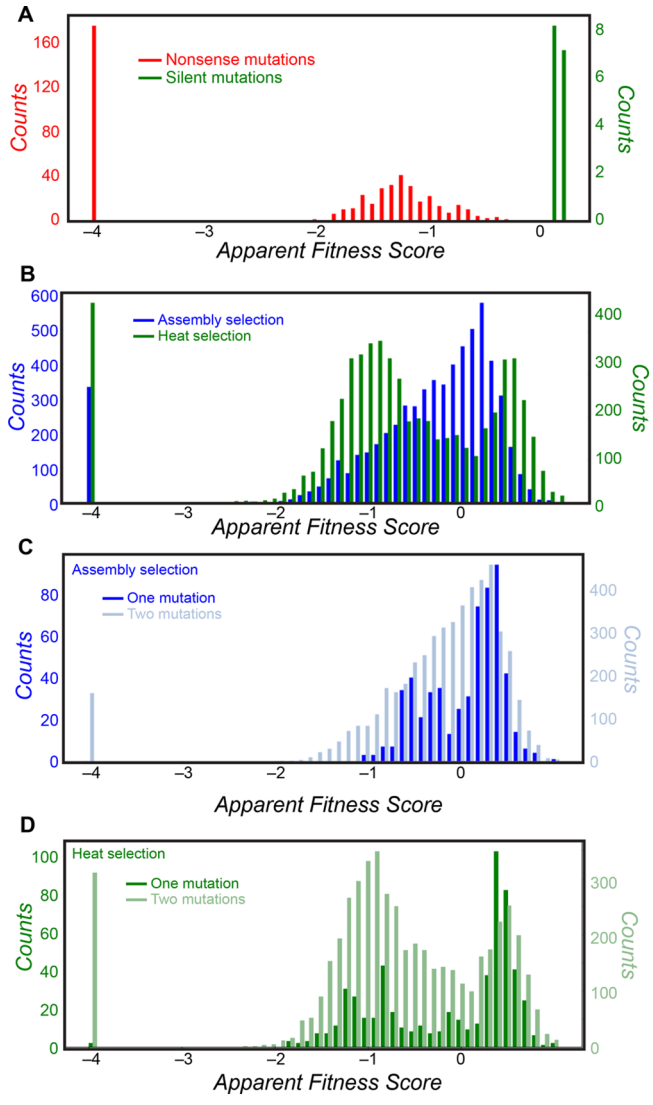


Figure 3. Assembly-selected apparent fitness score (AFS) abundances. (A) Nonsense and silent mutation AFS values separate into two nonoverlapping populations. (B) The heat-selected 2D AFL separates into bimodal populations, while the assembly selection 2D AFL does not. Single-amino acid mutations are on average less deleterious than two-amino acid mutations in the (C) assembly and (D) heat selections.

Selection for Thermal Stability Separates Wild-Type-like Variants from Thermally Compromised Variants. Thermostability is desirable in nearly all potential applications, and any variants used as vaccines, biomaterials, or drug delivery vehicles would likely require near-wild-type stability or better. To identify variants that are stable to high temperatures,

DOI: 10.1021/acs.biochem.8b00948
Biochemistry 2019, 58, 1527−1538

we subjected library members to a heat challenge of 50 °C for 10 min. We then compared the percent abundance of variants after this selection to that of the starting plasmid library, resulting in a heat-selected 2D AFL (Figure S2). Variants that assemble and are stable to 50 °C for 10 min result in positive scores and are colored blue.

We next compared the distribution of AFS values in the assembly-selected and heat-selected 2D AFLs. A histogram of all AFS values in the less stringent, assembly-selected case results in a broad distribution centered slightly above zero with a long negative tail (Figure 3B). In contrast, a histogram of the AFS values in the heat-selected case exhibits bimodality, suggesting that many assembly competent variants strongly alter the thermostability of the MS2 CP. Variants with a high AFS value in the heat-selected 2D AFL are likely to be best suited for applications such as drug delivery or imaging because of their uncompromised thermal stability. The bimodality of the heat-selected data set is also evident in the dark colors and obvious striped patterns in the landscape itself (Figure S2). For example, with few exceptions, mutations at G74 resulted in negative scores in the heat-selected 2D AFL, indicating that nearly every combination of mutations at this position resulted in undesirable VLP properties. Similar effects can be seen with V75. At V75, mutation to isoleucine and, in some cases, leucine resulted in thermostable VLPs, but few other mutations were tolerated. Taken together, the stringency of the thermal selection is useful for identifying which variants behave like wild-type VLPs.

We hypothesized that one and two missense mutations may have different average effects on protein fitness. In this data set, 570 single-amino acid mutants were scored while 5041 two-amino acid mutations were scored. When the AFS values of one and two missense mutations were compared, differences between these populations were apparent (Figure 3C). The histogram comparing these values clearly shows that one-amino acid mutations form a bimodal distribution, indicating the VLPs split into well-assembled and poorly assembled populations. In contrast, AFS values for two-amino acid mutations are distributed more evenly and are lower on average, suggesting that an additional mutation is more often detrimental to VLP assembly. These differences were exacerbated by the heat selection, in which two-amino acid mutations were clearly less tolerated than one-amino acid mutations (Figure 3D), though both populations exhibit bimodality. These results agree with literature reports that a second mutation, or an additional step away from wild type in a fitness landscape, often has an additive, negative effect on protein fitness.[16]

Selections for Increased Acid Sensitivity with Uncompromised Thermal Stability. In targeted drug delivery, a therapeutic cargo can be protected inside of a container until it is endocytosed into target cells; thus, selective release of drug cargo in the acidic environment of endosomes or lysosomes is potentially advantageous. In addition to its role in VLP assembly, the FG loop is critical for modulating the acid stability of the MS2 CP. Previously, we used 1D AFLs to show that mutations CP[T71H] and CP[E76C] increase VLP sensitivity to an acidic environment.[15] However, previous attempts to improve the acid sensitivity of CP[T71H] or CP[E76C] through rational design led to compromised long-term stability, suggesting that the properties of protein stability and acid sensitivity may be intertwined in a non-obvious fashion. As such, identifying variants with inversely correlated properties, high thermostability with reduced acid stability, is well-suited for protein engineering approaches.

We therefore additionally selected for variants with stability to high temperatures and increased acid sensitivity. These selections were performed on the assembled library, ensuring that variants are assembly competent. In addition, we specifically sought variants that behaved like the previously published CP[T71H] variant, which selectively precipitates under acidic conditions. We challenged the assembly-selected library of FG loop variants to pH 5 at 37 °C for 1 h, mimicking the conditions of the early endosome.[61] High-throughput sequencing was used to identify VLPs that were selectively absent following acidic pressure (Figure S3).

In particular, three variants, CP[T71H/E76P], CP[T71H/E76Q], and CP[T71H/E76T], exhibited increased acid sensitivity compared to that of the parent CP[T71H] variant (Figure 4A,B). In contrast, all instances of silent mutations that encode CP[WT] showed unchanged abundance following acidic pressure and increased relative abundance following thermal pressure (Figure S4). Interestingly, all three acid-sensitive variants contained the parent CP[T71H] mutation, combined with a second mutation at residue 76. Gratifyingly, at pH 7.2, all three variants exhibited a melting temperature of >50 °C, ranging from 52 to 65 °C (Figure 4C).

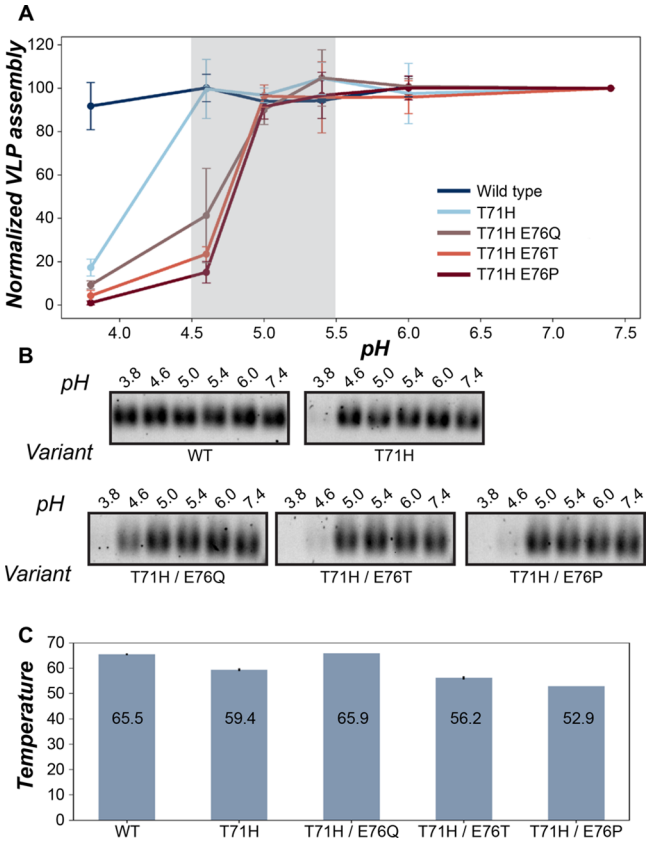


Figure 4. Validation of acid-sensitive, heat-stable variants. (A and B) Three new variants, CP[T71H/E76Q], CP[T71H/E76T], and CP[T71H/E76P], are more acid-sensitive than CP[T71H]. Capsid recoveries after acid challenges were quantified via native gel electrophoresis, followed by ImageJ densitometry analysis. Error bars represent three sample replicates. (C) All variants exhibit melting temperatures of >50 °C at pH 7.2.

DOI: 10.1021/acs.biochem.8b00948
Biochemistry 2019, 58, 1527−1538

Additionally, CP[T71H/E76P], the most acid-sensitive variant, tolerated an additional cysteine mutation in the interior cavity at position N87. This cysteine mutation has previously been used to load fluorophore or drug cargo into the interior of the MS2 CP.[62] The triple mutant CP[T71H/E76P/N87C] formed well-assembled VLPs and was readily modified by AlexaFluor-488 maleimide (Figure S5A); in contrast, the CP[T71H/E76P] double mutant lacking the introduced cysteine residue remained unmodified. Analysis by HPLC SEC confirmed that the AlexaFluor-488 fluorophore coelutes with CP[T71H/E76P/N87C] following modification (Figure S5B,C), consistent with the behavior of CP[T71H/N87C] and CP[N87C] (Figure S5D−G) and indicating that the modified VLPs are assembled. These useful two-amino acid variants display a narrow yet desirable combination of properties, achieved by combining a highly targeted library with multiple direct functional selections.

Quantifying FG Loop Pairwise Mutability Using Shannon Entropy. The previously published 1D AFL was used to quantify the mutability index (MI) of each position in the FG loop (Table 1). The region contains a range of mutabilities, from poorly mutable (G74 and V75) to highly mutable (T71, V72, G73, and E76). In this study, the Shannon entropy,[63] a calculation that measures diversity at a given position, was used to quantify the pairwise mutability at every combination of positions in both the assembly-selected and the heat-selected 2D AFL. As expected, positions with lower mutability, such as G74, decreased the pairwise mutability, even when combined with positions with high mutability, such as T71 or E76 (Figure S6A). Residues T71, V72, G73, and E76 are independently highly mutable with similar mutability indices, as determined by the 1D AFL. However, pairwise combinations containing E76 were less mutable than pairwise combinations of T71, V72, and G73, suggesting that multiple mutations carry an increased penalty at position E76.

Table 1

residue	mutability index
71	−0.04
72	−0.05
73	−0.01
74	−0.22
75	−0.23
76	−0.09

Differences in pairwise mutability were evaluated following heat selection (Figure S6B). In particular, the pairwise mutability of residue V75 decreased more than those of other positions, suggesting that combinations of mutations that include V75 may permit assembly but lead to a loss of thermostability. Only the pairwise combination of T71 and E76 inverted from mutable to immutable following thermal pressure, suggesting that mutations at both of these positions may be more deleterious to thermostability.

We next sought to predict the 2D mutability in this region using a convolutional neural network based on the 1D AFL (Figure S7A). The optimized convolutional neural network produced a 2D AFL in which the average mutability in each combination was comparable to the assembly-selected 2D AFL; however, individual combinations of amino acids were inconsistently predicted. Upon further analysis, we found that an additive 2D AFL, which was populated with the summed difference between the AFS value of both mutations and the average AFS_{WT} in the 1D AFL, outperformed the neural network in predicting the mutability of each combination (Figure S7B). We hypothesize that this performance discrepancy is due to the limited training data, which consisted of only 2580 mutants in the parent 1D AFL library. These efforts underscore the importance of continued experimental work to generate high-quality, multidimensional AFLs.

Epistasis Plays a Visible Role in Two-Amino Acid Mutability across the FG Loop. We sought to identify instances of negative and positive sign epistasis in the 2D AFL. Negative sign epistasis refers to combinations of mutations that do not permit assembly, even though each mutation is permitted individually. Conversely, positive sign epistasis refers to mutations that rescue nonassembling variants, resulting in assembled VLPs.

We identified variants for which the predicted 2D AFS value (calculated via the simple additive method described above) has a sign different from that of the measured assembly selected 2D AFS value. We quantified epistasis, E, as the difference between the predicted and experimental data sets (Figure 5A). While the majority of these variants have E scores closer to zero, a subset exhibits notably positive and negative sign epistasis (Data Set S3). The median E value is negative, consistent with previous studies of sign epistasis, which have shown that deleterious pairwise interactions are more common.[64]

Although much of the experimental landscape matches the predicted 2D AFL, we found that epistasis plays a surprising role in the two-amino acid mutability across many positions in this loop. In particular, positions T71 and E76, which are spatially adjacent but distant in sequence (Figure 2B,C), coevolve with a strong interdependence. Mutations at T71 change which mutations are permitted at E76, and vice versa, with regard to both charge and steric bulk. In addition, combinations of charged mutations are tightly regulated by epistatic effects across the entire loop, in which some evolutionary paths are more available than others.

As examples of this, both positive sign epistasis and negative sign epistasis were observed at combinations of mutations at positions T71 and E76, in the assembly- and heat-selected 2D AFL (Figure 5B). While distant in sequence, these positions are structurally close in both the A/C and B conformations. In the A/C conformation (quasi-6-fold axis), T71 and E76 are on adjacent β-sheets, while in the B conformation (5-fold axis), the side chain of E76 hydrogen bonds with the backbone of T71 (Figure 2C).[65,66]

In the 1D AFL, a single negative charge at position T71 is not permitted. We previously hypothesized that this is likely due to repulsion with the nearby negative charge at E76. This hypothesis is supported by the 2D AFL results. T71E can be rescued by a charge inversion at E76, and CP[T71E/E76K], CP[T71E/E76R], and CP[T71E/E76H] are assembly competent and enriched following the thermal challenge (Figure 5C). More strikingly, T71D variants were rescued by almost any mutation at E76, with the notable exception of negatively charged residues (D and E) and structurally disruptive residues (G and P) (Figure 5D).

Visualization in Chimera yielded useful insight into the origin of this pattern. In CP[WT], E76 hydrogen bonds with both Q40 and the backbone of T71 (Figure S8A) in the B form monomer structure. When E76 is inverted from a hydrogen acceptor to a hydrogen donor (R and K), hydrogen

DOI: 10.1021/acs.biochem.8b00948
Biochemistry 2019, 58, 1527−1538

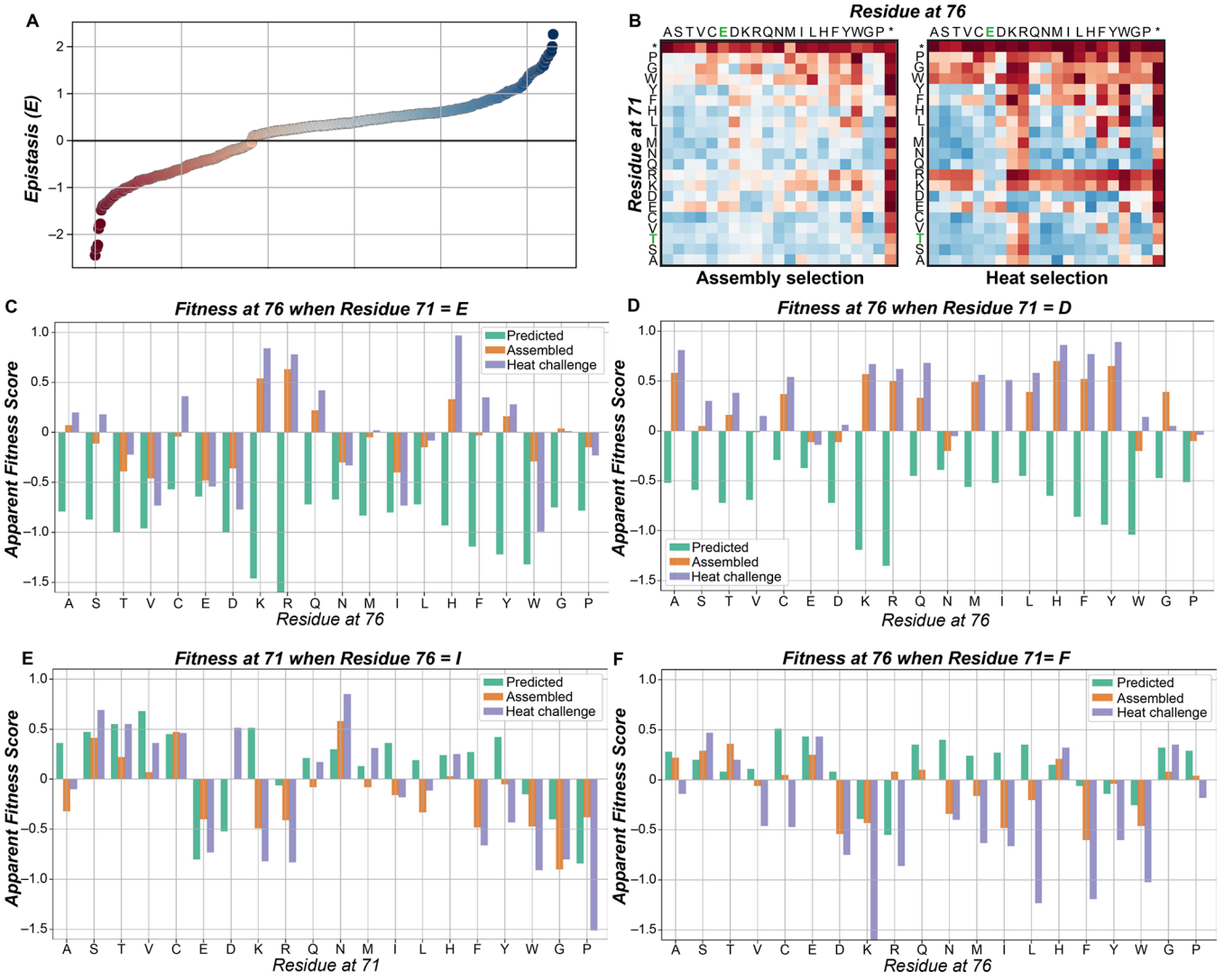


Figure 5. Positive and negative epistasis in the FG loop. (A) *E*, a measure of epistasis, shows both negative and positive effects, though the overall trend is toward negative epistasis. (B) Residues 71 and 76 show significant positive and negative epistasis, as shown for both the assembly and heat selections. Examples of (C and D) predominantly positive epistasis and (E and F) negative epistasis are shown. In these graphs, 2D AFS values predicted from the 1D AFL data are colored green and compared to the measured assembly selected (orange) and heat-selected (purple) AFS values. Differences between the predicted and measured scores indicate regions of epistatic interactions.

bonding with Q40 could be preserved, as glutamine contains both an acceptor and a donor. However, hydrogen bonding between E76 and the backbone of T71 is likely not preserved without significant backbone rearrangement. In addition, in the case of CP[T71E/E76R], the mutated side chains are oriented in opposite directions, again indicating that a new salt bridge is likely not formed without backbone rearrangement in this region (Figure S8B). Finally, *in silico* mutation to CP[E76R] results in clashes with Q70, further supporting the idea that backbone rearrangement of this flexible loop is likely. Taken together, we anticipate that doubly charged mutants at T71 and E76 are engaging in backbone rearrangement in the B form, thus restoring a mimic of the native hydrogen bonding pattern or permitting a salt bridge between variant side chains.

Instances of negative sign epistasis were also observed between positions T71 and E76. In the 1D AFL, each position independently permits hydrophobic mutations. For example, CP[E76I] is assembly competent; however, when coupled with an additional mutation of isoleucine, leucine, phenylalanine, or tyrosine at T71—all mutations that are tolerated individually—the VLP is no longer assembly competent (Figure 5E). Similarly, CP[T71F] is permitted in the 1D AFL, but when combined with additional hydrophobic mutations at position E76, the VLP no longer permits assembly (Figure 5F).

These trends suggest a steric constraint between these two residues, where enough space exists for one but not multiple bulky amino acids. Visualization in Chimera led us to hypothesize that the steric constraint is driven by the structure of the A/C form monomers rather than the B form monomers. In the C form, and to a lesser extent in the A form, clashes between the two bulky residues were apparent following *in silico* mutation to CP[T71F/E76I] (Figure S8C), while the B form allowed mutation without producing clashes. Given the secondary structure in this region, we hypothesize that these clashes are not readily resolved, leading to assembly incompetent VLPs.

Across the FG loop, combinations of oppositely charged mutations have varying, and often striking, effects (Figure S9A). The negative charge at T71 is rescued by the positive charge at V72, G73, or E76. Similar, though subtler, effects are

DOI: 10.1021/acs.biochem.8b00948
Biochemistry 2019, 58, 1527−1538

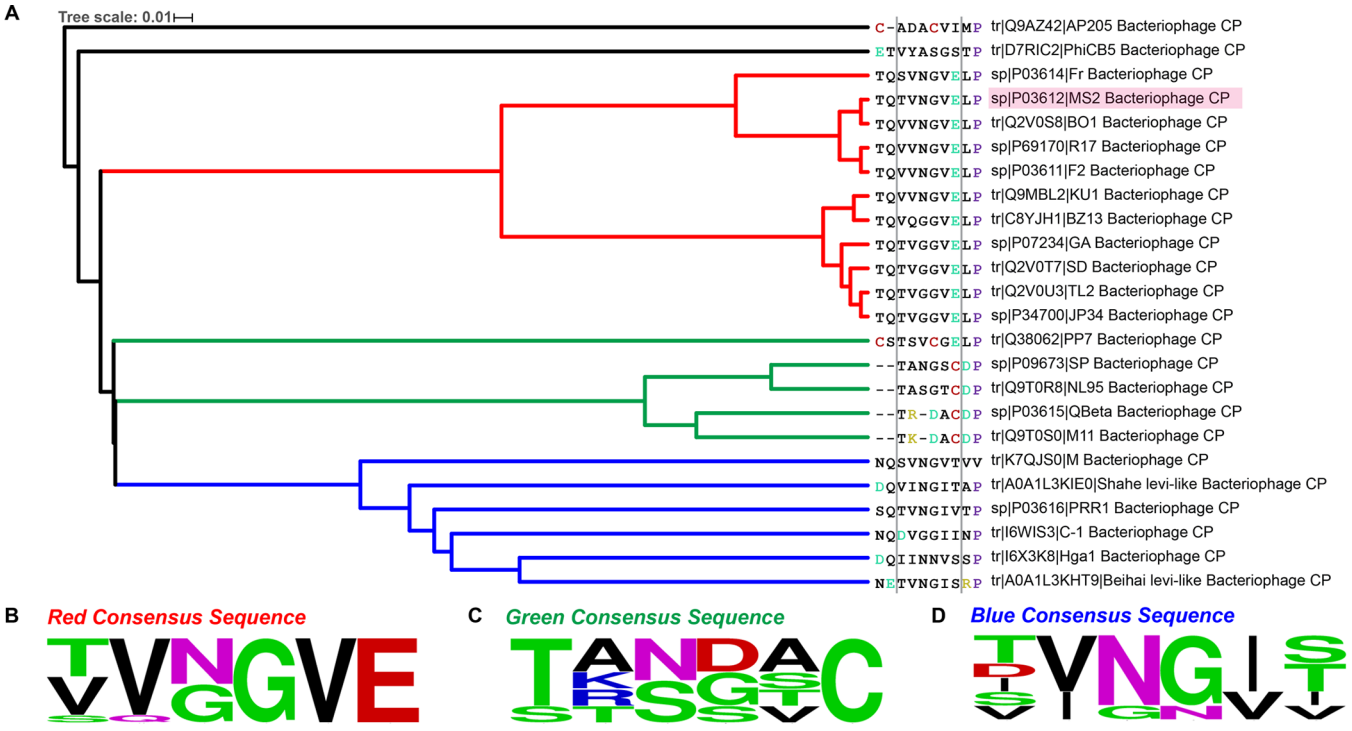


Figure 6. Phylogenetic tree of RNA bacteriophage coat proteins. (A) Twenty-four coat proteins separated into three distinct clades, colored red, blue, and green. The gray lines indicate the regions that are analogous to the FG loop in bacteriophage MS2. (B–D) These clades exhibit distinct consensus sequences.

seen for the negative charge at V72 with the corresponding positive charge at T71 or G73, and combinations of oppositely charged mutations at these positions lead to well-assembled, thermostable VLPs. Similarly, charge inversion at E76, which is not tolerated in the 1D AFL, is rescued in several instances with negative charge elsewhere in the loop.

A different trend is observed with charged mutations at position V72 in combination with position G74 or V75, both poorly mutable positions in general. While assembly of VLPs with charge at V72 is rescued by a second, oppositely charged mutation at G74 or V75, these mutations lead to thermally unstable VLPs that do not remain assembled through the heat challenge. Similar trends are observed when the charge at G74 and V75 is combined with an oppositely charged mutation at E76. While these trends are challenging to explain in full, we visualized the hydrogen bond network and local environment of the FG loop in the A/C and B forms to gain more insight (Figure S8D,E). Within a 5 Å region, the A/C form largely interacts with other regions of nearby FG loop, albeit a region wider than that evaluated in this study. In contrast, the FG loop of the B form is spatially near a wide range of residues, including an adjacent loop region. In addition, several key hydrogen bonds appear to maintain the correct geometric shape in this loop. As such, we hypothesize that the strong effect of charge on the FG loop is driven by the hydrogen bonding network in the B form monomer.

Other instances of sign epistasis were less intuitive. We observed non-obvious positive epistasis between residues G73 and V75. A tyrosine mutation at V75 is not tolerated when G73 is the wild-type residue. However, surprisingly, positive charge or hydrophobic mutations, including K, R, F, L, M, and Y, rescue the assembly and thermostability of V75Y (Figure S9B). How positive charge rescues a mutation to tyrosine is unclear. One explanation could be that the positive charge forces a new interaction with the negatively charged E76 residue, contorting the loop enough to allow the bulky V75Y mutation. An alternative possibility is a new cation–π interaction between V75Y and K or R or π-stacking with F and Y.

The only two paired residues that show no evidence of epistasis are G74 and V75, the least mutable residues in this region. Of the 400 possible combinations of mutations at G74 and V75, only CP[G74G/V75V], CP[G74G/V75I], and CP[G74G/V75L] form thermostable VLPs.

Taken together, these detailed analyses work toward defining the design rules for successful VLP formation of the MS2 CP, adding depth to our understanding of self-assembly.

Phylogenetic Analysis Suggests a Relationship between Epistasis and an Evolutionary Path. Epistatic analysis produced several compelling design rules governing the mutability—and, potentially, evolvability—of the FG loop. We sought to compare these rules against sequences of homologous coat proteins from related bacteriophages. A phylogenetic tree from sequenced RNA phages was generated, and these 24 bacteriophage coat proteins separated into three distinct clades (Figure 6A). This tree is similar to previously published phylogenetic analyses of ssRNA phages.[67] Consistent with the design rules generated by epistatic analysis, most homologues have zero or one negatively charged residues in this loop. Additional negative residues are frequently accompanied by positive residues, and residues corresponding to positions 71 and 76 are not both large, hydrophobic amino acids.

Bacteriophages closely related to the MS2 CP show a high degree of sequence similarity in the region corresponding to the FG loop. No charged residues are observed other than E76. Additionally, residues T71, V72, and G73 exhibit the most sequence diversity, consistent with our mutability analyses. In

DOI: 10.1021/acs.biochem.8b00948
Biochemistry 2019, 58, 1527–1538

contrast, a clade of structurally related coat proteins that includes the well-studied Qβ bacteriophage[26] shows a divergent consensus sequence (Figure 6B). In this case, the negative charge at E76 is replaced with a cysteine residue. Interestingly, this cysteine is known to participate in intersubunit disulfide bonding. In addition, exposure to DTT decreases the melting temperature of the Qβ CP by >40 °C, indicating that this disulfide bond is critical for thermostability.[68] Within this group, we also see more sequence divergence at residues 71−75, including several charged residues at positions 72 and 74. The third clade, which is most distant from the MS2 CP, shows high divergence from the MS CP sequence, with the exception of the conserved G74 residue. In this group, most sequences have only one negative charge, primarily at a position corresponding to residue 71 in the MS2 CP.

From these analyses, we hypothesize that the observed effect of charge on the MS2 CP FG loop arises in the absence of disulfide bonds present in the clade containing Qβ. Without intersubunit disulfide bonds, the MS2 CP likely relies on the hydrogen bond network in the B form monomer, including the intrasubunit hydrogen bond between residues T71 and E76, to maintain the correct geometric shape, and disrupting this interaction with additional charge, charge inversion, or other mutations can result in poorly formed VLPs. However, compensating with a similar hydrogen bond through an additional mutation or balancing additional charge with an opposite charge can be restorative.

CONCLUSION

VLPs are an excellent model system for studying the effects of epistasis on protein assembly, because of their genetic simplicity, high yield, and intrinsic genotype-to-phenotype link. We generated and characterized a 6615-member library of one- and two-amino acid mutations in the highly mutable FG loop of the MS2 CP. The library was subjected to multiple selections, initially for capsid assembly, followed by thermal and pH stability. Negative or positive epistasis was identified and characterized. In particular, two-amino acid mutations involving charged residues and steric bulk coevolved in unexpected ways. Our epistatic analysis was used to generate a set of design rules for this loop, which were compared to consensus sequences for related coat protein clades in a phylogenetic analysis. This study represents the first quantitative measure of epistasis in a self-assembling nanoparticle, and the generated design rules will inform future efforts to tailor and engineer viruslike particles in a variety of non-native contexts.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.biochem.8b00948.

Data Set 1: Primers for library generation, high-throughput sequencing, and individual variant cloning (XLSX)

Data Set 2: Complete dataset for assembly- and heat-selected 2D AFLs (XLSX)

Data Set 3: List of candidate two amino acid variants exhibiting epistasis (XLSX)

Additional figures, primers, complete 2D AFL scores, and epistatic scores (PDF)

AUTHOR INFORMATION

Corresponding Authors

*E-mail: ercek@northwestern.edu.
*E-mail: mbfrancis@berkeley.edu.

ORCID

Matthew B. Francis: 0000-0003-2837-2538
Danielle Tullman-Ercek: 0000-0001-6734-480X

Author Contributions

E.C.H., M.J.L., M.B.F., and D.T.-E. conceived this project. E.C.H. and S.A.R. performed experiments for this project, and E.C.H., A.H.F., and M.J.L. analyzed the apparent fitness landscapes. E.C.H., M.B.F., and D.T.-E. wrote the manuscript. All of the authors reviewed and contributed to the manuscript.

Funding

This work was funded by the Army Research Office (W911NF-15-1-0144 and W911NF-16-1-0169) and the BASF CARA program at the University of California, Berkeley. E.C.H. was supported under by the Department of Defense, Air Force Office of Scientific Research, National Defense Science and Engineering Graduate (NDSEG) Fellowship (32 CFR 168a).

Notes

The authors declare no competing financial interest.
Apparent fitness landscape and epistatic data for the MS2 CP are available in the Supporting Information, and additional information is available from the authors by request.
All in-house code is available from the authors upon request.

ACKNOWLEDGMENTS

The authors thank Daniel Brauer, Varun Tolani, Steven Strutt, Helen Hobbs, and Charlotte Nixon for helpful discussions and instrumentation to conduct these analyses. The sequencing was performed by the DNA Technologies and Expression Analysis Cores at the University of California Davis Genome Center, supported by National Institutes of Health Shared Instrumentation Grant S10 OD010786.

REFERENCES

(1) Sailer, Z. R., and Harms, M. J. (2017) High-Order Epistasis Shapes Evolutionary Trajectories. *PLoS Comput. Biol. 13* (5), e1005541.
(2) Steinberg, B., and Ostermeier, M. (2016) Environmental Changes Bridge Evolutionary Valleys. *Sci. Adv. 2* (1), e1500921−e1500921.
(3) Miton, C. M., and Tokuriki, N. (2016) How Mutational Epistasis Impairs Predictability in Protein Evolution and Design. *Protein Sci. 25*, 1260−1272.
(4) Gong, L. I., and Bloom, J. D. (2014) Epistatically Interacting Substitutions Are Enriched during Adaptive Protein Evolution. *PLoS Genet. 10* (5), e1004328.
(5) Gupta, A., and Adami, C. (2016) Strong Selection Significantly Increases Epistatic Interactions in the Long-Term Evolution of a Protein. *PLoS Genet. 12* (3), e1005960.
(6) Wu, N. C., Du, Y., Le, S., Young, A. P., Zhang, T.-H., Wang, Y., Zhou, J., Yoshizawa, J. M., Dong, L., Li, X., Wu, T.-T., and Sun, R. (2016) Coupling High-Throughput Genetics with Phylogenetic Information Reveals an Epistatic Interaction on the Influenza A Virus M Segment. *BMC Genomics 17* (1), 46.

DOI: 10.1021/acs.biochem.8b00948
Biochemistry 2019, 58, 1527−1538

(7) Weinreich, D. M., Delaney, N. F., DePristo, M. A., and Hartl, D. L. (2006) Darwinian Evolution Can Follow Only Very Few Mutational Paths to Fitter Proteins. *Science 312* (5770), 111−114.
(8) Tenaillon, O., Rodríguez-Verdugo, A., Gaut, R. L., McDonald, P., Bennett, A. F., Long, A. D., and Gaut, B. S. (2012) The Molecular Diverstiy of Adaptive Convergence. *Science (Washington, DC, U. S.) 335* (6067), 457−462.
(9) Kryazhimskiy, S., Dushoff, J., Bazykin, G. A., and Plotkin, J. B. (2011) Prevalence of Epistasis in the Evolution of Influenza A Surface Proteins. *PLoS Genet. 7* (2), e1001301.
(10) Firnberg, E., Labonte, J. W., Gray, J. J., and Ostermeier, M. (2014) A Comprehensive, High-Resolution Map of a Gene's Fitness Landscape. *Mol. Biol. Evol. 31* (6), 1581−1592.
(11) Wrenbeck, E. E., Azouz, L. R., and Whitehead, T. A. (2017) Single-Mutation Fitness Landscapes for an Enzyme on Multiple Substrates Reveal Specificity Is Globally Encoded. *Nat. Commun. 8*, 15695.
(12) Fowler, D. M., Araya, C. L., Fleishman, S. J., Kellogg, E. H., Stephany, J. J., Baker, D., and Fields, S. (2010) High-Resolution Mapping of Protein Sequence-Function Relationships. *Nat. Methods 7* (9), 741−746.
(13) Tinberg, C. E., Khare, S. D., Dou, J., Doyle, L., Nelson, J. W., Schena, A., Jankowski, W., Kalodimos, C. G., Johnsson, K., Stoddard, B. L., et al. (2013) Computational Design of Ligand-Binding Proteins with High Affinity and Selectivity. *Nature 501* (7466), 212−216.
(14) Roscoe, B. P., Thayer, K. M., Zeldovich, K. B., Fushman, D., and Bolon, D. N. A. (2013) Analyses of the Effects of All Ubiquitin Point Mutants on Yeast Growth Rate. *J. Mol. Biol. 425* (8), 1363−1377.
(15) Hartman, E. C., Jakobson, C. M., Favor, A. H., Lobba, M. J., Álvarez-Benedicto, E., Francis, M. B., and Tullman-Ercek, D. (2018) Quantitative Characterization of All Single Amino Acid Variants of a Viral Capsid-Based Drug Delivery Vehicle. *Nat. Commun. 9* (1), 1385.
(16) Sarkisyan, K. S., Bolotin, D. A., Meer, M. V., Usmanova, D. R., Mishin, A. S., Sharonov, G. V., Ivankov, D. N., Bozhanova, N. G., Baranov, M. S., Soylemez, O., et al. (2016) Local Fitness Landscape of the Green Fluorescent Protein. *Nature 533* (7603), 397−401.
(17) Melamed, D., Young, D. L., Gamble, C. E., Miller, C. R., and Fields, S. (2013) Deep Mutational Scanning of an RRM Domain of the Saccharomyces Cerevisiae Poly(A)-Binding Protein. *RNA 19* (11), 1537−1551.
(18) Jacquier, H., Birgy, A., Le Nagard, H., Mechulam, Y., Schmitt, E., Glodt, J., Bercot, B., Petit, E., Poulain, J., Barnaud, G., et al. (2013) Capturing the Mutational Landscape of the Beta-Lactamase TEM-1. *Proc. Natl. Acad. Sci. U. S. A. 110* (32), 13067−13072.
(19) Bershtein, S., Segal, M., Bekerman, R., Tokuriki, N., and Tawfik, D. S. (2006) Robustness−Epistasis Link Shapes the Fitness Landscape of a Randomly Drifting Protein. *Nature 444* (7121), 929−932.
(20) Bank, C., Hietpas, R. T., Jensen, J. D., and Bolon, D. N. A. (2015) A Systematic Survey of an Intragenic Epistatic Landscape. *Mol. Biol. Evol. 32*, 229−238.
(21) Olson, C. A., Wu, N. C., and Sun, R. (2014) A Comprehensive Biophysical Description of Pairwise Epistasis throughout an Entire Protein Domain. *Curr. Biol. 24* (22), 2643−2651.
(22) Li, C., Qian, W., Maclean, C. J., and Zhang, J. (2016) The Fitness Landscape of a TRNA Gene. *Science (Washington, DC, U. S.) 352* (6287), 837−840.
(23) Glasgow, J., and Tullman-Ercek, D. (2014) Production and Applications of Engineered Viral Capsids. *Appl. Microbiol. Biotechnol. 98* (13), 5847−5858.
(24) Slininger Lee, M., and Tullman-Ercek, D. (2017) Practical Considerations for the Encapsulation of Multi-Enzyme Cargos within the Bacterial Microcompartment for Metabolic Engineering. *Curr. Opin. Syst. Biol. 5*, 16−22.
(25) Kanaan, N. M., Sellnow, R. C., Boye, S. L., Coberly, B., Bennett, A., Agbandje-McKenna, M., Sortwell, C. E., Hauswirth, W. W., Boye, S. E., and Manfredsson, F. P. (2017) Rationally Engineered AAV Capsids Improve Transduction and Volumetric Spread in the CNS. *Mol. Ther.–Nucleic Acids 8*, 184−197.
(26) Fiedler, J. D., Higginson, C., Hovlid, M. L., Kislukhin, A. A., Castillejos, A., Manzenrieder, F., Campbell, M. G., Voss, N. R., Potter, C. S., Carragher, B., et al. (2012) Engineered Mutations Change the Structure and Stability of a Virus-like Particle. *Biomacromolecules 13* (8), 2339−2348.
(27) Anishchenko, I., Ovchinnikov, S., Kamisetty, H., and Baker, D. (2017) Origins of Coevolution between Residues Distant in Protein 3D Structures. *Proc. Natl. Acad. Sci. U. S. A. 114* (34), 9122−9127.
(28) Sutter, M., Greber, B., Aussignargues, C., and Kerfeld, C. A. (2017) Assembly Principles and Structure of a 6.5-MDa Bacterial Microcompartment Shell. *Science (Washington, DC, U. S.) 356* (6344), 1293−1297.
(29) Ashley, C. E., Carnes, E. C., Phillips, G. K., Durfee, P. N., Buley, M. D., Lino, C. A., Padilla, D. P., Phillips, B., Carter, M. B., Willman, C. L., et al. (2011) Cell-Specific Delivery of Diverse Cargos by Bacteriophage MS2 Virus-like Particles. *ACS Nano 5* (7), 5729−5745.
(30) Galaway, F. A., and Stockley, P. G. (2013) MS2 Viruslike Particles: A Robust, Semisynthetic Targeted Drug Delivery Platform. *Mol. Pharmaceutics 10* (1), 59−68.
(31) ElSohly, A. M., Netirojjanakul, C., Aanei, I. L., Jager, A., Bendall, S. C., Farkas, M. E., Nolan, G. P., and Francis, M. B. (2015) Synthetically Modified Viral Capsids as Versatile Carriers for Use in Antibody-Based Cell Targeting. *Bioconjugate Chem. 26* (8), 1590−1596.
(32) Aanei, I. L., Elsohly, A. M., Farkas, M. E., Netirojjanakul, C., Regan, M., Taylor Murphy, S., O'Neil, J. P., Seo, Y., and Francis, M. B. (2016) Biodistribution of Antibody-MS2 Viral Capsid Conjugates in Breast Cancer Models. *Mol. Pharmaceutics 13* (11), 3764−3772.
(33) Farkas, M. E., Aanei, I. L., Behrens, C. R., Tong, G. J., Murphy, S. T., O'Neil, J. P., and Francis, M. B. (2013) PET Imaging and Biodistribution of Chemically Modified Bacteriophage MS2. *Mol. Pharmaceutics 10*, 69−76.
(34) Jeong, K., Netirojjanakul, C., Munch, H. K., Sun, J., Finbloom, J. A., Wemmer, D. E., Pines, A., and Francis, M. B. (2016) Targeted Molecular Imaging of Cancer Cells Using MS2-Based 129Xe NMR. *Bioconjugate Chem. 27* (8), 1796−1801.
(35) Peabody, D. S., Manifold-Wheeler, B., Medford, A., Jordan, S. K., do Carmo Caldeira, J., and Chackerian, B. (2008) Immunogenic Display of Diverse Peptides on Virus-like Particles of RNA Phage MS2. *J. Mol. Biol. 380* (1), 252−263.
(36) Tumban, E., Peabody, J., Tyler, M., Peabody, D. S., and Chackerian, B. (2012) VLPs Displaying a Single L2 Epitope Induce Broadly Cross-Neutralizing Antibodies against Human Papillomavirus. *PLoS One 7* (11), e49751.
(37) Capehart, S. L., ElSohly, A. M., Obermeyer, A. C., and Francis, M. B. (2014) Bioconjugation of Gold Nanoparticles through the Oxidative Coupling of Ortho -Aminophenols and Anilines. *Bioconjugate Chem. 25* (10), 1888−1892.
(38) Capehart, S. L., Coyle, M. P., Glasgow, J. E., and Francis, M. B. (2013) Controlled Integration of Gold Nanoparticles and Organic Fluorophores Using Synthetically Modified MS2 Viral Capsids. *J. Am. Chem. Soc. 135* (8), 3011−3016.
(39) Hietpas, R. T., Jensen, J. D., and Bolon, D. N. A. (2011) Experimental Illumination of a Fitness Landscape. *Proc. Natl. Acad. Sci. U. S. A. 108* (19), 7896−7901.
(40) Engler, C., Gruetzner, R., Kandzia, R., and Marillonnet, S. (2009) Golden Gate Shuffling: A One-Pot DNA Shuffling Method Based on Type IIs Restriction Enzymes. *PLoS One 4* (5), e5553.
(41) Pinto, F., Thapper, A., Sontheim, W., and Lindblad, P. (2009) Analysis of Current and Alternative Phenol Based RNA Extraction Methodologies for Cyanobacteria. *BMC Mol. Biol. 10* (1), 79.
(42) Wang, W., and Malcolm, B. A. (1999) Two-Stage PCR Protocol Allowing Introduction of Multiple Mutations, Deletions and Insertions Using QuikChange Site-Directed Mutagenesis. *BioTechniques 26* (4), 680−682.

DOI: 10.1021/acs.biochem.8b00948
Biochemistry 2019, 58, 1527−1538

(43) Bolger, A. M., Lohse, M., and Usadel, B. (2014) Trimmomatic: A Flexible Trimmer for Illumina Sequence Data. *Bioinformatics 30* (15), 2114−2120.
(44) Magoč, T., and Salzberg, S. L. (2011) FLASH: Fast Length Adjustment of Short Reads to Improve Genome Assemblies. *Bioinformatics 27* (21), 2957−2963.
(45) Li, H., and Durbin, R. (2009) Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform. *Bioinformatics 25* (14), 1754−1760.
(46) Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009) The Sequence Alignment/Map Format and SAMtools. *Bioinformatics 25* (16), 2078−2079.
(47) Shin, H.-C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., and Summers, R. M. (2016) Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. *IEEE Trans. Med. Imaging 35* (5), 1285−1298.
(48) Pontius, J., Richelle, J., and Wodak, S. J. (1996) Deviations from Standard Atomic Volumes as a Quality Measure for Protein Crystal Structures. *J. Mol. Biol. 264* (1), 121−136.
(49) Sandberg, M., Eriksson, L., Jonsson, J., Sjöström, M., and Wold, S. (1998) New Chemical Descriptors Relevant for the Design of Biologically Active Peptides. A Multivariate Characterization of 87 Amino Acids. *J. Med. Chem. 41* (14), 2481−2491.
(50) Krigbaum, W. R., and Komoriya, A. (1979) Local Interactions as a Structure Determinant for Protein Molecules II. *Biochim. Biophys. Acta, Protein Struct. 576* (1), 204−228.
(51) Fauchère, J.-L., Charton, M., Kier, L. B., Verloop, A., and Pliska, V. (1988) Amino Acid Side Chain Parameters for Correlation Studies in Biology and Pharmacology. *Int. J. Pept. Protein Res. 32* (4), 269−278.
(52) Fasman, G. D. (1989) *Practical Handbook of Biochemistry and Molecular Biology*, Wiley.
(53) Letunic, I., and Bork, P. (2016) Interactive Tree of Life (ITOL) v3: An Online Tool for the Display and Annotation of Phylogenetic and Other Trees. *Nucleic Acids Res. 44* (W1), W242−W245.
(54) Crooks, G. E., Hon, G., Chandonia, J. M., and Brenner, S. E. (2004) WebLogo: A Sequence Logo Generator. *Genome Res. 14* (6), 1188−1190.
(55) Peabody, D. S. (1997) Subunit Fusion Confers Tolerance to Peptide Insertions in a Virus Coat Protein. *Arch. Biochem. Biophys. 347* (1), 85−92.
(56) Caldeira, J. C., and Peabody, D. S. (2011) Thermal Stability of RNA Phage Virus-like Particles Displaying Foreign Peptides. *J. Nanobiotechnol. 9* (1), 22.
(57) Caspar, D. L., and Klug, A. (1962) Physical Principles in the Construction of Regular Viruses. *Cold Spring Harbor Symp. Quant. Biol. 27*, 1−24.
(58) Rolfsson, O., Toropova, K., Ranson, N. A., and Stockley, P. G. (2010) Mutually-Induced Conformational Switching of RNA and Coat Protein Underpins Efficient Assembly of a Viral Capsid. *J. Mol. Biol. 401* (2), 309−322.
(59) Ni, C. Z., Syed, R., Kodandapani, R., Wickersham, J., Peabody, D. S., and Ely, K. R. (1995) Crystal Structure of the MS2 Coat Protein Dimer: Implications for RNA Binding and Virus Assembly. *Structure 3* (3), 255−263.
(60) Glasgow, J. E., Capehart, S. L., Francis, M. B., and Tullman-Ercek, D. (2012) Osmolyte-Mediated Encapsulation of Proteins inside MS2 Viral Capsids. *ACS Nano 6* (10), 8658−8664.
(61) Meng, F., Cheng, R., Deng, C., and Zhong, Z. (2012) Intracellular Drug Release Nanosystems. *Mater. Today 15*, 436−442.
(62) Stephanopoulos, N., Tong, G. J., Hsiao, S. C., and Francis, M. B. (2010) Dual-Surface Modified Virus Capsids for Targeted Delivery of Photodynamic Agents to Cancer Cells. *ACS Nano 4* (10), 6014−6020.
(63) Stewart, J. J., Lee, C. Y., Ibrahim, S., Watts, P., Shlomchik, M., Weigert, M., and Litwin, S. (1997) A Shannon Entropy Analysis of Immunoglobulin and T Cell Receptor. *Mol. Immunol. 34* (15), 1067−1082.
(64) Starr, T. N., and Thornton, J. W. (2016) Epistasis in Protein Evolution. *Protein Sci. 25* (7), 1204−1218.
(65) Golmohammadi, R., Valegård, K., Fridborg, K., and Liljas, L. (1993) The Refined Structure of Bacteriophage MS2 at 2·8 Å Resolution. *J. Mol. Biol. 234* (3), 620−639.
(66) Dai, X., Li, Z., Lai, M., Shu, S., Du, Y., Zhou, Z. H., and Sun, R. (2017) In Situ Structures of the Genome and Genome-Delivery Apparatus in a Single-Stranded RNA Virus. *Nature 541* (7635), 112−116.
(67) Kannoly, S., Shao, Y., and Wang, I.-N. (2012) Rethinking the Evolution of Single-Stranded RNA (SsRNA) Bacteriophages Based on Genomic Sequences and Characterizations of Two R-Plasmid-Dependent SsRNA Phages, C-1 and Hgal1. *J. Bacteriol. 194* (18), 5073−5079.
(68) Ashcroft, A. E., Lago, H., Macedo, J. M. B., Horn, W. T., Stonehouse, N. J., and Stockley, P. G. (2005) Engineering Thermal Stability in RNA Phage Capsids via Disulphide Bonds. *J. Nanosci. Nanotechnol. 5* (12), 2034−2041.

DOI: 10.1021/acs.biochem.8b00948
Biochemistry 2019, 58, 1527−1538