Published in final edited form as:
Nat Biotechnol. 2014 June ; 32(6): 562–568. doi:10.1038/nbt.2880.

Quantitative analysis of RNA-protein interactions on a massively parallel array for mapping biophysical and evolutionary landscapes

Jason D. Buenrostro[1,3], **Lauren M. Chircus**[1,2], **Carlos L. Araya**[1], **Curtis J. Layton**[1], **Howard Y. Chang**[3], **Michael P. Snyder**[1], and **William J. Greenleaf**[1,*]

[1]Department of Genetics, Stanford University School of Medicine, Stanford, CA, USA

[2]Department of Chemical and Systems Biology, Stanford University School of Medicine, Stanford, CA, USA

[3]Program in Epithelial Biology and the Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, CA, USA

Abstract

RNA-protein interactions drive fundamental biological processes and are targets for molecular engineering, yet quantitative and comprehensive understanding of the sequence determinants of affinity remains limited. Here we repurpose a high-throughput sequencing instrument to quantitatively measure binding and dissociation of MS2 coat protein to $>10^7$ RNA targets generated on a flow-cell surface by *in situ* transcription and inter-molecular tethering of RNA to DNA. We decompose the binding energy contributions from primary and secondary RNA structure, finding that differences in affinity are often driven by sequence-specific changes in association rates. By analyzing the biophysical constraints and modeling mutational paths describing the molecular evolution of MS2 from low- to high-affinity hairpins, we quantify widespread molecular epistasis, and a long-hypothesized structure-dependent preference for G:U base pairs over C:A intermediates in evolutionary trajectories. Our results suggest that quantitative analysis of RNA on a massively parallel array (RNAMaP) relationships across molecular variants.

RNA-protein interactions drive a wide variety of critical biological processes from gene expression[1] to viral assembly[2]. Up to 10% of the eukaryotic proteome is estimated to bind RNA[3], and recent work has begun to uncover a web of RNA-protein interactions[4-6] that can

[*]Correspondence to: wjg@stanford.edu.

Author Contributions
W.J.G, J.D.B, and C.J.L. conceived of the method. J.D.B developed and performed the RNA display protocol. L.M.C designed and performed the protein purification and binding assays. J.D.B wrote the image analysis algorithm with input from W.J.G and C.J.L. C.L.A. developed and implemented the structural (epistatic), functional (modeling, kinetic), and evolutionary analyses. All authors interpreted the data and wrote the manuscript. W.J.G supervised all aspects of this work.

Competing financial interests
The authors declare no competing financial interests.

Accession codes
Sequence data are available at the NCBI Sequence Read Archive (SRA) under the accession number SRP040276.

control gene expression through splicing, RNA localization, and other post-transcriptional processes. Protein interactions with long noncoding RNAs also play a role in epigenetic state changes during differentiation[7], perhaps through "scaffolding" chromatin remodelers[8,9]. Furthermore, RNA-protein interactions have proven powerful tools in synthetic biology, allowing gene expression control through post-transcriptional regulation[10,11].

A biophysical understanding of the nucleic-acid sequence determinants of RNA-protein interactions lags behind our growing realization of their biological importance. Unlike double-stranded DNA, RNA substrates demonstrate diverse intramolecular interactions—including, mismatched base bulges, stem loops, pseudo knots, g-quartets, divalent cation interactions and non-canonical base pairs—that determine three-dimensional RNA structure[12-15] and set the landscape for interactions with RNA-binding proteins (RBPs)[16]. The combinatorial nature of RNA sequence and intramolecular interactions, coupled with the relative paucity of data produced from current biophysical methods has precluded a high-resolution, predictive understanding of both the sequence dependence of affinity and the resulting evolutionary constraints imposed by these requirements. Because the relationship between sequence and binding is often opaque, little is understood regarding the evolutionary constraints on these RNA structures, making bioinformatic identification of functional RNAs difficult[17].

Current methods for investigating the sequence dependence of RNA-protein interactions include medium-throughput microfluidic methods[18] and high-throughput methods coupling affinity-based selection with high-throughput DNA sequencing or array hybridization[19] and recently have been used to generate a catalogue of RNA binding motifs[20].

Although powerful, selection and sequencing methods bias results towards high-activity variants and do not directly and quantitatively measure the biophysical parameters that underlie biological function[21]. Recently, methods have been developed to quantitatively measure catalysis[22,23], however, no such high-throughput methods exists for determining binding parameters k_{on}, k_{off} and K_d for RNA-protein interactions.

The technological innovations that have propelled the high-throughput sequencing revolution provide the foundations for massively parallel, fluorescence-based observations over a large variety of nucleic acid structures immobilized on a surface[24-27]. Recent work characterizing DNA-protein interactions[27] has demonstrated the utility of these instruments for high-throughput binding affinity assays across large DNA sequence space. In this work, we have leveraged the Illumina DNA sequencing platform, an instrument that integrates solid-phase molecular biology, fluidics and high-throughput TIRF imaging for massively parallel DNA sequencing[28], to create a platform for direct, ultra-high throughput measurement of RNA-protein interactions. In addition, we have developed quantitative image analysis tools for large-scale analysis of these data, and demonstrate measurement of both equilibrium binding constants and dissociation kinetics. We apply these methods to the MS2 coat protein[2,29-33], a system with widespread applications in affinity purification[34], RNA imaging[35] and synthetic biology[10,11]. This approach enables quantitative measurement of binding and dissociation of a protein to $>10^7$ RNA targets generated directly on the flow cell surface, providing massive biophysical datasets enabling predictive models for affinity tuning, decomposition of binding energies between primary and secondary structures and quantitative analysis of evolutionary trajectories across sequence space.

Results

A high-throughput RNA array for quantitative measurements

To generate a library of RNA targets, we first made an Illumina sequencing library containing an *E. coli* RNA polymerase (RNAP) initiation and stall sequence and a region coding for diverse sequence variants of the MS2 RNA hairpin synthesized using doped oligonucleotides (Fig. 1a,b, Supplementary Fig. 1, Supplementary Table 1). To ensure multiple measurements of each RNA variant and reduce sequencing error[36], we introduced single-molecule barcodes 5' of the RNAP initiation sequence. The barcoding strategy serves to identify individual molecules within a population by uniquely tagging each molecule using a barcode. We then diluted the amplification reaction such that approximately 8×10^5 molecules were amplified in the reaction, which created a “bottleneck” in the population of barcoded molecular variants. This procedure allowed for each barcoded molecular species to be sequenced a median of 15 times per sequencing lane, allowing for multiple redundant measurements across the flow cell (**Supplementary Fig. 2**). The sequencing process converted individual molecules within the library to ~1 μm diameter clusters of ~1,000 clonal DNA molecules on the flow cell surface[28], and provided the sequence and position of the DNA templates across the 2D array.

Following sequencing, we removed the sequenced DNA strand, and regenerated double stranded DNA (dsDNA) using DNA polymerase to extend a biotinylated primer. We then saturated the flow cell with streptavidin to create a terminal biotin-streptavidin roadblock on these dsDNA fragments. To synthesize RNA, we adapted methods from single-molecule investigations[37] designed to generate a single RNA per DNA template. First, we initiated *E. coli* RNA polymerase holoenzyme (RNAP) in CTP-starved conditions, which allows RNAP to generate 26 bases of RNA (the footprint of RNAP) before stalling at the first guanine on the DNA template strand. Second, we washed excess RNA polymerase from solution and introduced all four nucleotides, allowing RNAP to transcribe the variable region and stall at the biotin-streptavidin roadblock. This procedure results in transcribed RNA tethered to its parent DNA via RNA polymerase (**Fig. 1a, Supplementary Fig. 3**). The resulting RNA array contained 1.2×10^7 distinct RNA features comprising 1.48×10^5 unique sequences in a single sequencing lane.

Quantitative binding and dissociation measurements

To measure binding energies, we flowed MS2 coat protein fluorescently labeled with SNAP14 Surface 549 over the RNA array, and imaged bound MS2 protein at equilibrium using total internal reflection fluorescence (TIRF) at 10 increasing concentrations. After the final measurement, we perfused 1.8 μM unlabeled MS2 protein and recorded the fluorescence decay caused by dissociation (**Fig. 1c, Supplementary Movie 1**). The high-concentration of unlabeled MS2 protein blocks other binding sites on the array, preventing re-binding of fluorescently labeled MS2. To quantify bound MS2 protein, we developed image analysis tools that cross-correlate cluster centers from sequencing data to acquired images and fit the observed binding in each cluster to a 2D Gaussian (Supplementary Fig. 4-5, software is available as Supplementary Data). Using this approach, we quantified the fluorescence signal for each cluster in 6,240 images representing 120 tiles imaged in two fluorescence color channels across 11 equilibrium MS2 concentrations and 15 dissociation time points. Fluorescence signals from single clusters fit canonical dissociation (**Fig. 1d, Supplementary Fig. 6**) and binding curves (**Fig. 1e, f, Supplementary Fig. 7**), yielding binding energy estimates in excellent agreement with published measurements ($R = 0.94$, slope = 1.08, **Fig. 1g**) and *in vitro* binding assays ($R = 0.92$, slope = 0.76, **Supplementary Fig. 8**).

We calculated off rates (k_{off}) for 3,029 sequences and dissociation constants (K_{d}) for 129,248 sequences, encompassing 57 single (100%), 1,539 double (100%), and 24,181 triple (92.4%) mutants (**Fig. 2a, b,** for data see **Supplementary Table 2-3**, for error estimation and quality control **Supplementary Fig. 9-10**). To investigate how sequence variation in the RNA hairpin impacts MS2 binding, we examined differential binding energies for all single-mutants compared to the consensus sequence ($-\Delta\Delta G_{\mathrm{consensus}}=0\ k_{\mathrm{B}}T$). The average binding energy change from all possible single-base changes at each position reveals a sensitivity to mutation throughout the hairpin that complements the effects of mutating individual residues on the binding surface of MS2 to alanine[38] (**Fig. 2c and Supplementary Fig. 11**). Specifically, we observe high mutation sensitivity at base-paired positions near the loop and at specific single-stranded positions, suggesting significant primary sequence and secondary structure requirements for RNA recognition.

Affinity partitioned between primary and secondary structure

To comprehensively examine these primary and secondary structure effects on binding, we calculated the $-\Delta\Delta G$ of all double-mutants (**Fig. 2d**). We observed high positive epistasis in a population of “compensating mutants”, suggesting that these pairs of mutations preserve hairpin structure and maintain high binding affinities (**Fig. 2e**). We also observed negative epistasis in non-compensating mutants near the base of the stem, potentially due to cooperative effects on hairpin destabilization in these regions. Reciprocal mapping of positive epistasis signatures (≥1 s.d.) allowed *de novo* reconstruction of the bound hairpin structure, identifying base-paired, loop, and bulge positions (**Supplementary Fig. 12a**) demonstrating the feasibility of reconstructing molecular RNA structures from large-scale sequence-function data.

We modeled the contributions of base-specificity (primary structure) and base-pairing (secondary structure) to binding energy at each position in the hairpin with a linear regression model from a set of 121 training sequences. This model provides two free parameters for each unpaired base accounting for primary sequence changes in the form of transitions or transversions. For each pair of interacting bases, the model provides a total of six free parameters–one for transition and transversion of each base in the pair (four parameters) as well as one parameter to account for disruption owing to the loss of base-pairing and one parameter representing possible non-canonical base-pairing interactions. These parameters were optimized jointly, in order to identify (via regression) the energetic contributions of primary sequence changes (i.e. transitions or transversions that occur while holding secondary structure constant) and secondary structure changes (i.e. inferred energetic consequences of secondary structure disruptions or formation of non-canonical bases in isolation from primary sequence perturbations). To quantify the sensitivity for non-canonical base-pairing at positions in the hairpin stem, we trained the model eight separate times (once for each possible non18 canonical pairing) with one free parameter representing the energetic cost of the respective non-canonical pairing. This re-fitting analysis allowed the model to incorporate a different energetic penalty for having non-canonical base pairs at a specific position instead of the energetic penalty for a full loss of base-pairing. In this analysis, G:U base pairs caused substantially less disruption to the binding energy than other non-canonical base pairs (**Fig. 3a**), consistent with the formation of a wobble base pair at G:U positions that allows partial rescue of the secondary structure[12,39]. Our final model, which incorporated a free parameter for G:U non-canonical base pairs, captured 92% of the variance in binding energy of the training set (**Supplementary Fig. 12b**) and predicted the binding energy of second and third mutations for variants with mutations in both paired and unpaired positions with correlation coefficients R=0.94 and R=0.83, respectively (**Fig. 3b**).

The model fit parameters allowed quantitative decomposition of primary and secondary determinants of affinity across the RNA structure (**Fig. 3c, d**). Energetic penalties for disrupting base-pairing increase with proximity to the loop, while non-canonical G:U base pairs cause substantially less energetic disruption at the –8:–3 and –11:–1 positions. Altering the primary sequence at –10A (bulge) and –4A (loop), residues that interact with the Lys61 binding pocket on alternate halves of the dimer[31], confers energetic costs that exceed disrupting the hairpin structure at any single base pair. We also observed important roles for the –7A and –5C residues, consistent with stacking interactions at these positions[40]. Altering the primary sequence on the 5’ side of the hairpin confers a greater energetic penalty compared with altering the 3’ side, which we speculate results from direct interactions with MS2 on the 5’ side[38].

Association rate contributes to changes in binding energies

We sought to quantify how changes in association and dissociation rates contribute to measured $-\Delta\Delta G$ values for all mutants with measurable kinetic data. We calculated the energetic contributions to $-\Delta\Delta G$ from changes in dissociation rates $\left[-log\left(k_{off}^{\text{mutant}}/k_{off}^{\text{consensus}}\right)\underset{=}{def}\Delta log\left(k_{off}^{\text{mutant}}\right)\right]$, and inferred the contribution from changes in association rates, $\left[log\left(k_{on}^{\text{mutant}}/k_{on}^{\text{consensus}}\right)\underset{=}{def}\Delta log\left(k_{on}^{\text{mutant}}\right)\right]$. Because $\Delta\log(k_{\text{off}})$ + $\Delta\log(k_{\text{on}}) = -\Delta\Delta G$, we treated these parameters as pseudo-energies. Using this decomposition, we examined the fractional contribution of change in dissociation rates to $-\Delta\Delta G$ across single and double mutants (**Fig. 4a**). At the base of the hairpin, only a small fraction of $-\Delta\Delta G$ measurements are explained by dissociation rate changes. This small effect suggests that mutations at these positions modulate association rates, possibly by causing fraying of the hairpin and/or allowing competition with alternate RNA structures, thereby reducing the per-collision probability of productive binding (see Supplementary Discussion). This interpretation is reinforced by examining $\Delta\log(k_{\text{off}})$ and $\Delta\log(k_{\text{on}})$ in this region (**Fig. 4b, c**). Dissociation rates change little while inferred association rates remain similar to that of the consensus sequence only for structures that maintain base-pairing through compensating mutations. Across all measured variants, we observe a significant population of structures with $-\Delta\Delta G$ driven by association rates (**Fig. 4d**; $P < 2.2 \times 10^{-16}$, Wilcoxon signed rank test, $\mu = 0.5$). These results suggest the kinetic drivers of observed affinity changes are position-specific and often operate through modulating association rates, likely by changing hairpin stability.

Analysis of quantitative evolutionary landscapes

We sought to understand how biophysical properties shape RNA sequence evolution towards higher binding affinity by examining the prevalence of epistasis, or differential mutational path probabilities caused by non-additive affinity gains, in molecular evolution—a question of intense debate[41,42]. Following previous work[43,44], we reconstructed 1,997 complete sets of mutational paths (tesseracts) describing the probability of evolving through permutations of four mutations from 1,597 low-affinity to 127 high-affinity hairpins. We modeled the probability of mutation, or the traversal from a source to a target node, as the effective probability of MS2 binding to the target over all sequences within one mutation of the source in the tesseract. Mutations can arise in any order, resulting in N=4!=24 distinct paths through which mutations may be sequentially acquired (**Fig. 5a**), with path probabilities defined as the product of probabilities for each mutational step. This model allows us to examine how molecular evolution towards higher affinity could proceed in an RNA-protein interaction, a question separate from the *in vivo* evolutionary landscape of MS2 sequences where the relationship between affinity and cellular fitness, as well as pleiotropic roles of this sequence in the MS2 genome, define the contours of the fitness landscape.

We examined evolutionary constraint (E_{AUC}), defined as the area under the curve of the cumulative probability of rank-ordered paths, in each set[43] (**Fig. 5a**). The data from 47,928 mutational paths revealed strong constraint in evolution towards higher affinity, with 81% of path probability contained within the top 30% of mutational paths (**Fig. 5b**). The observed evolutionary constraint exceeds that expected from a non-epistatic landscape accounting for measurement errors (null model), or from a model that assumes a random distribution of affinities. These results indicate that distributions of affinity effects in mutational paths are highly structured (**Fig. 5c**), consistent with widespread intramolecular epistasis in evolutionary phase space[41,43-45].

The sum of the mutational path probabilities (E_{SUM}) captures the probability of reaching a given high-affinity sequence from a given low-affinity sequence. We observed a non-uniform distribution of both evolutionary probability (E_{SUM}) and constraint (E_{AUC}) from tesseracts involving mutations at different residues in the hairpin structure (**Fig. 5d, e**) implying that biophysical properties impose strong, systematic, structure-dependent effects on evolutionary trajectories.

By modeling evolutionary sequence preference of ribosomal RNA, Rousset *et al*. observed that trajectories transitioning from A:U to G:C base pairs preferentially traverse G:U versus A:C intermediates and hypothesized this non-canonical base-pairing as a general mechanism for maintaining RNA-protein contacts in evolution[46]. Data from 696 tesseracts containing both G:U and A:C intermediates reveal differential preferences for paths traversing G:U intermediates across the hairpin stem (**Fig. 5f, g**), providing evidence that biophysical properties underling the preference for G:U intermediates derive not from universal properties of secondary structure, but from the details of the RNA-protein interaction. With the exception of one position (**Supplementary Figure 13**), we observe no strong differences between the path probabilities of G:A and U:C intermediates for U:A to G:C transitions highlighting the contextual dependencies of these path probabilities.

Discussion

Using *in situ* transcription and inter-molecular tethering of RNA to DNA, we have converted a high-throughput DNA sequencing flow cell into an RNA array for quantitatively measuring both binding kinetics and thermodynamics at an unprecedented scale. Using this quantitative deep mutational profiling approach we report, to our knowledge, the largest collection of binding affinities and kinetic constants for an intermolecular interaction. Using this dataset, we addressed long-standing biophysical questions, including (i) the relative contributions of primary and secondary structure elements to binding energy, (ii) the sequence-dependent kinetic contributions to observed affinities, (iii) the prevalence of evolutionary epistasis and (iv) the context-dependence of preference for G:U intermediates in secondary structure.

Our predictive model for RNA-protein affinity across thousands of point mutations provides a map for quantitative tuning of both the association rate and the equilibrium constants of this RNA-protein interaction. We anticipate this resource of sequence variants will enable affinity tuning of MS2-based RNA sensors enabling new applications in synthetic biology. Additionally, these data provide quantitation of the effect of primary sequence, secondary structure and non-canonical base-pairing, creating a valuable framework for understanding the design and evolution of new RNA aptamers.

We hypothesize that inferred changes in on-rates are due to destabilization of the RNA hairpin formation or competition with alternate secondary structure, reducing the number of productive binding collisions[47] (Supplementary Discussion). These observations suggest the data provided here may also provide a rich resource for modeling the RNA hairpin stability and alternate structure formation. While this is an area of inquiry beyond the focus of this work, the potential for formation of alternate structures and the effects of local sequence on native folding of RNA are well suited for study using this platform, as the RNA transcripts are synthesized by *E. coli* RNAP and folded co-transcriptionally, closely approximating synthesis conditions *in vivo*.

We observe that evolutionary landscapes of RNA-protein interactions are highly constrained, further supporting a major role for intramolecular epistasis in shaping evolutionary trajectories and providing insight into complexities of both natural and human-directed evolutionary methods for generating high-affinity ligands. Our analysis provides a quantitative mapping of G:U bias in evolutionary intermediates that has been previously observed46. However, our observation complicates the simple assumption that G:U bias is simply a function of regions of RNA that form secondary structure and interact strongly with RNA. By observing a lack of G:U/C:A bias at the –9 base pair adjacent to the adenine bulge, we note that this preference is dependent on the context and the specifics of the secondary structure in this region.

We anticipate this RNA-MaP methodology will be a powerful addition to selection- and sequencing-based methods. In addition, the technique might provide quantitative information on RNA libraries generated by systematic enrichment of ligands by exponential enrichment (SELEX), allowing affinity tuning for the design of biological parts. Although SELEX methods often begin with large libraries (~10^{14}) and produce a small number of selected molecules, our RNA array methodology allows quantitative characterization of a much larger library subset (~10^{5}), opening the door to a detailed understanding of the sequence-specific rules driving acquisition of affinity in the selection process. Alternatively, our approach might be coupled to sequenced *in vivo* RNA immunoprecipitation libraries[48,49] and used to directly quantify molecular affinities on *in vitro* generated RNA, providing measurements of interactions in well-defined conditions. The multicolor imaging capabilities of the sequencer enables measurement of more complex biological interactions such as cooperativity between differentially labeled binding partners or RNA structure inference via fluorescence resonance energy transfer (FRET). In addition, the sequencing platform is capable of generating DNA clusters >1kb[50], enabling transcription of long RNAs and allowing investigations of long non-coding RNAs and catalytic ribozymes (see Supplementary Discussion for possible limitations). In short, we believe future application of RNA-MaP to diverse RNA-protein and RNA-RNA interactions promises to enable quantitative prediction and engineering of binding affinities and functional RNA molecules, as well as the identification and understanding of evolutionary sequence constraints based on underlying biophysical parameters.

Online Methods

Library Design and Construction

To generate a high density RNA array, we designed a custom DNA library containing a barcode, *E. coli* RNA polymerase (RNAP) promoter, RNAP stall sequence, constant region, and degenerate MS2 hairpin sequence (**Supplementary Fig. 1**). The reverse compliment of the region containing the stall sequence, constant region, and MS2 hairpin were synthesized by the Stanford Protein and Nucleic Acid Facility. Degenerate bases were introduced into the MS2 hairpin region using hand-mixed bases containing 88% of the consensus base and 4% of each non-consensus base. This degeneracy ratio was chosen to maximize the total number of variants represented on the RNA array as well as the fractional representation of triple mutants. The RNAP promoter, barcode, and Illumina sequencing primers and adapters were subsequently added by PCR.

Library Bottlenecking and Amplification

The sequencing library was then bottlenecked to ensure multiple measurements of each RNA variant (**Supplementary Fig. 2**). To quantify the amount of starting material, we used a prequantified commercially available PhiX library (Illumina) as a concentration standard.

PhiX library was diluted to 50 pM then diluted 1:2 seven times in 10 mM Tris pH 8 + 0.01% Tween20 to create a dilution series ranging from 50 pM to 0.39 pM. For each concentration of diluted PhiX and for the assembled MS2 hairpin library, 1 μL of library was added to a qPCR mix containing 1x NEBnext PCR Mix, 1.25μM oligos C and D, and 0.6x Sybr Green. qPCR was carried out for 40 cycles, and the C_t values for each PhiX dilution and library sample were obtained. For PhiX, the concentration of each sample was plotted against the C_t value and was fit to a line. Using the resulting equation, we related C_t to concentration and calculated the concentration of the MS2 hairpin library. We then diluted the MS2 hairpin library to approximately 30.6 fM (~9.2×10^5 molecules) in 50 μL of the same PCR mix and amplified the library to approximately 30 nM (21 cycles).

Sequencing Amplified Libraries

Libraries were sequenced on an Illumina GAIIx to a cluster density of 1.23×10^7 clusters per lane. The libraries were sequenced in 2 steps using the standard single-end sequencing protocol. First, 15 cycles were used to read the barcode, and then 27 cycles were used to read the variable hairpin region. Reading the random 15 bp barcode first improved sequencing quality (data not shown) due to higher sequence diversity of the first 15 cycles of sequencing. Sequencing was done by ELIM Biopharmaceuticals (Hayward, CA).

MS2 Coat Protein Purification

The MS2-dlFG mutant[30] of the MS2 Coat Protein was cloned into a custom expression vector containing an *N*-terminal FLAG and SNAPtag (NEB) and a *C*-terminal 6xHis tag (https://benchling.com/s/oYAOq4). The construct was transformed into BL21(DE3) cells (NEB), and starter cultures of transformed cells were grown overnight in a rotator at 37°C in 10 mL LB with 50 μg/mL kanamycin. 500 mL LB with 50 μg/mL kanamycin was inoculated with 10 mL overnight starter culture and grown shaking at 37°C for 2.5 hours. SNAPtag-MS2 expression was induced with 0.5 mM IPTG for 5 hours at 22°C, and then cells were collected by centrifugation at 4000 rpm for 15 minutes at 4°C. Cell pellets were frozen at −20°C overnight. MS2 protein was purified using the Qiagen Ni-NTA Fast Start Kit. To maximize purity, twice the suggested amount of cell pellet was used, cell lysis was extended to one hour, the flow through was reapplied to the column 5 times, and the column was washed 2 times with 8 mL wash buffer. Purified protein was dialyzed 1:1,000,000 into 100 mM Ultrapure Tris-HCl, pH 8.0 (Invitrogen), 150 mM NaCl, and 1 mM DTT using Slide-A-Lyzer 7000MWCO dialysis cassettes (Thermo). Protein was quantified by A280 absorption on a NanoDrop and Coomassie Plus Protein Assay (Thermo). Attempts to purify an MS2-dlFG fused to tagRFP in place of the SNAPtag via the same protocol resulted in protein aggregation in culture and on the sequencing chip (data not shown).

Labeling MS2 Coat Protein with SNAPtag Substrate

5 μM purified SNAPtag-MS2 was labeled with SNAP-Surface 549 fluor (NEB) at 37°C for 30 minutes in 50 mM Tris pH 8.0, 100 mM NaCl, 0.1% Tween 20, 1 mM DTT, and 10 μM SNAPSurface 549. Excess SNAP-Surface 549 was removed using Zeba Spin Desalting Columns (Thermo) equilibrated with TMK Buffer (100 mM Tris-HCl pH 8.0, 80 mM KCl, 10 mM $MgCl_2$, 1 mM DTT).

RNA Labeling and Filter Binding Assays

RNA variants were obtained from IDT and the Stanford Protein and Nucleic Acid Facility. RNAs were diluted to 5 μM in 10 μL end labeling reactions of 1x T4 PNK buffer with 10 units PNK (NEB) and 5 μCi gamma-ATP. Excess gamma-ATP was removed with the Zymo Oligo Clean and Concentrator kit. Approximately 20 pM labeled RNA was then incubated with varying concentrations of MS2 ranging from 0 to 8100 nM in TMKG buffer (TMK buffer, 10% glycerol, 100 μ g/mL BSA) for 1.75 hours at room temperature. The MS2/RNA mixtures were then filtered through a nitrocellulose membrane (GE) followed by a positively charged nylon membrane (GE) then Whatman paper on a dot-blot apparatus (Bio-Rad) using the house vacuum (**Supplementary Fig. 8a**). Membranes were allowed to air dry before exposure to a phosphor screen for 12-96 hours. Phosphor screens were scanned on a Typhoon and the signal from each dot was quantified in ImageJ. Fraction bound (f_{bound}) was determined for each filtered MS2/RNA mixture as the signal on the nitrocellulose ($signal_{\text{nitrocellulose}}$) (which binds protein and therefore MS2-RNA complexes) over the total signal on both the nitrocellulose and the positively charged nylon ($signal_{\text{+Nylon}}$) (which binds free RNA).

$$f_{bound}=\frac{signal_{nitrocellulose}}{signal_{nitrocellulose}+signal_{+Nylon}}$$

The concentration of protein (C) versus fraction bound was fit to a bimolecular binding curve in MATLAB for each of three replicates to find the K_{d}. (Fit parameters f_{max}=maximal fraction bound and f_{max}=minimum fraction bound.)

$$f_{bound}=\frac{f_{max}}{1+\frac{K_d}{C}}+f_{min}$$

Modifications to the Illumina Genome Analyzer IIx (GAIIx)

To improve the optics and allow for equilibrium measurements on an Illumina sequencer, we modified the sequencer in several ways. First, we exchanged the standard Illumina fluorescence filter to a filter optimized for SNAP-Surface 549 fluorescence emission (Semrock FF01-562/40-25). Second, we eliminated unwanted wash steps after imaging and during the "safe state" mode by changing the default SCS files. C:\Illumina\SCS2.10\DataCollection\bin\Config\HCMConfig.xml was modified to: <SafeStatePump Solution="4" AspirationRate="250" DispenseRate="2500" Volume="0" />, and C:\Illumina\SCS2.10\DataCollection\bin\Config\ImageCyclePump.xml was modified to <ImageCyclePump On="false" AutoDispense="false">. We also shortened all the fluidics lines of the GAIIx and the associated paired-end module.

Generation of the RNA array

All subsequent steps were performed on the modified GAIIx using GAIIx software running custom fluidics and imaging scripts. After sequencing, dsDNA clusters on the Illumina flow cell were denatured using 0.1 N NaOH. Following denaturing, we observed residual fluorescence from the sequencing reaction (**Supplementary Fig. 3a,b**). Therefore, we incorporated an additional cleavage step (100 mM Tris, 125 mM NaCl, 100 mM TCEP, 50 mM sodium ascorbate, and 0.05% Tween 20) (**Supplementary Fig. 3c**). Following cleavage, we annealed a 5’ biotinylated primer to the 3’ sequencing adaptor and resynthesized dsDNA using Klenow DNA polymerase (1× NEB buffer 2, 250 μM dNTP mix, 0.1 units/μl NEB Klenow, 0.01% Tween-20) incubated for 30 minutes at 37°C. We then flowed in 100 nM RNase free streptavidin to bind to the 5’ biotinylated primer and passivized with a 500 nM biotin wash. To block all potential ssDNA, we annealed an unlabeled oligo complementary to the constant stall sequence. We then incubated the dsDNA with a transcription initiation mix containing sigma saturated RNAP and three nucleotides at 2.5 μM (1x T7A1 reaction buffer [20 mM Tris, 20 mM NaCl, 7 mM $MgCl_2$, 0.1 mM EDTA, 0.1 % BME, 0.02 mg/ml BSA, 1.5% glycerol], 2.5 μM e ach A TP, GTP a nd U TP, 0.015 mg/ml RNAP [Sigma-saturated holoenzyme from Epicentre] and 0.01% tween-20) for 30 minutes at 37°C. In this buffer, RNAP initiates onto dsDNA clusters and stalls at the first cytosine, generating 26 bases of RNA. Stalled RNAP covers the initiation site to inhibit multiple RNAPs from initiating on the same DNA molecule. Excess RNAP was washed from solution with 1x T7A1 reaction buffer plus 2.5 μM each ATP, GTP and UTP. Finally, 10 mM NTPs (ATP, CTP, GTP, and UTP) in 1x T7A1 reaction buffer were added for 30 minutes at 37°C to allow transcription to proceed. After transcription, RNAP remained stalled at the 5’ biotinstreptavidin roadblock, generating a stable RNAP mediated DNA-RNA tether (**Fig. 1a**).

MS2 Binding and Dissociation Experiments on the RNA Array

To assay total synthesized RNA, we annealed an Alexa Fluor 647 labeled DNA oligo onto the stall sequence that was present in all clusters (**Supplementary Fig. 1, 3**). Based on cluster fluorescence intensities, we observed an RNA synthesis efficiency of approximately 30-40%. We also annealed an unlabeled MS2_3’block oligo to the constant region between the hairpin and the RNAP footprint in order to prevent alternate secondary structures. Following annealing, we assayed binding by introducing SNAP-Surface 549-MS2 (TMK buffer, 100 μg/mL BSA and 10 μg/mL yeast tRNAs) to the flow cell at 3x increasing concentrations starting at 0.046 nM and ending at 900 nM for a total of 10 binding images. For each measurement, we waited 1 hour to reach equilibrium. Following binding at 900 nM MS2, we observed dissociation by introducing 1.8 μM unlabeled MS2 and continually imaging the 120 tiles of the flow cell.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by NIH R01-HG004361 (to H.Y.C.); H.Y.C. is an Early Career Scientist of the Howard Hughes Medical Institute. JDB and LMC acknowledge support from NSF Graduate Research Fellowships. JDB also acknowledges support from NIH training grant RC4NS073015. LMC acknowledges support from NIH training grant GM067586-09. We thank D. Herschlag for feedback and advice throughout the methods development, and G. Sherlock and D. Herschlag for critical readings of earlier versions of this manuscript. We thank R. Landick for discussions regarding the design of our synthetic DNA library and K. Bajaj for discussions regarding quantitation of cluster fluorescence.

References and Notes

1. Keene JD. RNA regulons: coordination of post-transcriptional events. Nature Reviews Genetics. 2007; 8:533–543.
2. Carey J, Cameron V, De Haseth PL, Uhlenbeck OC. Sequence-specific interaction of R17 coat protein with its ribonucleic acid binding site. Biochemistry. 1983; 22:2601–2610. [PubMed: 6347247]
3. Tsvetanova NG, Klass DM, Salzman J, Brown PO. Proteome-Wide Search Reveals Unexpected RNA-Binding Proteins in Saccharomyces cerevisiae. PLoS ONE. 2010; 5:e12671. [PubMed: 20844764]
4. Scherrer T, Mittal N, Janga SC, Gerber AP. A Screen for RNA-Binding Proteins in Yeast Indicates Dual Functions for Many Enzymes. PLoS ONE. 2010; 5:e15499. [PubMed: 21124907]
5. Butter F, Scheibe M, Morl M, Mann M. Unbiased RNA-protein interaction screen by quantitative proteomics. Proceedings of the National Academy of Sciences. 2009; 106:10626–10631.
6. Castello A, et al. Insights into RNA Biology from an Atlas of Mammalian mRNA-Binding Proteins. Cell. 2012; 149:1393–1406. [PubMed: 22658674]
7. Wang KC, et al. A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. Nature. 2011; 472:120–124. [PubMed: 21423168]
8. Tsai MC, et al. Long Noncoding RNA as Modular Scaffold of Histone Modification Complexes. Science. 2010; 329:689–693. [PubMed: 20616235]
9. Guttman M, Rinn JL. Modular regulatory principles of large non-coding RNAs. Nature. 2012; 482:339–346. [PubMed: 22337053]
10. Culler SJ, Hoff KG, Smolke CD. Reprogramming Cellular Behavior with RNA Controllers Responsive to Endogenous Proteins. Science. 2010; 330:1251–1255. [PubMed: 21109673]
11. Ausländer S, Ausländer D, Müller M, Wieland M, Fussenegger M. Programmable single-cell mammalian biocomputers. Nature. 2012; 487:123–127. [PubMed: 22722847]
12. SantaLucia J, Turner DH. Measuring the thermodynamics of RNA secondary structure formation. Biopolymers. 1997; 44:309–319. [PubMed: 9591481]
13. Kertesz M, et al. Genome-wide measurement of RNA secondary structure in yeast. Nature. 2010; 467:103–107. [PubMed: 20811459]
14. Ding Y, et al. In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. Nature. 2013; 505:696–700. [PubMed: 24270811]
15. Rouskin S, Zubradt M, Washietl S, Kellis M, Weissman J. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. Nature. 2013; 505:701–705. [PubMed: 24336214]
16. Ban N, Nissen P, Hansen J, Moore PB, Steitz TA. The Complete Atomic Structure of the Large Ribosomal Subunit at 2.4 Å Resolution. Science. 2000; 289:905–920. [PubMed: 10937989]
17. Wan Y, Kertesz M, Spitale RC, Segal E, Chang HY. Understanding the transcriptome through RNA structure. Nature Reviews Genetics. 2011; 12:641–655.
18. Martin L, et al. Systematic reconstruction of RNA functional motifs with high-throughput microfluidics. Nat Meth. 2012; 9:1192–1194.
19. Ray D, et al. Rapid and systematic analysis of the RNA recognition specificities of RNAbinding proteins. Nat. Biotechnol. 2009; 27:667–670. [PubMed: 19561594]
20. Ray D, et al. A compendium of RNA-binding motifs for decoding gene regulation. Nature. 2013; 499:172–177. [PubMed: 23846655]
21. Araya CL, et al. A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. Proceedings of the NationalAcademy of Sciences. 2012; 109:16858–16863.
22. Pitt JN, Ferre-D'Amare AR. Rapid Construction of Empirical RNA Fitness Landscapes. Science. 2010; 330:376–379. [PubMed: 20947767]
23. Guenther U-P, et al. Hidden specificity in an apparently nonspecific RNA-binding protein. Nature. 2013 doi:10.1038/nature12543.

24. Matzas M, et al. High-fidelity gene synthesis by retrieval of sequence-verified DNA identified using high-throughput pyrosequencing. Nat. Biotechnol. 2010; 28:1291–1294. [PubMed: 21113166]

25. Myllykangas S, Buenrostro JD, Natsoulis G, Bell JM, Ji HP. Efficient targeted resequencing of human germline and cancer genomes by oligonucleotide-selective sequencing. Nat. Biotechnol. 2011; 29:1024–1027. [PubMed: 22020387]

26. Uemura S, et al. Real-time tRNA transit on single translating ribosomes at codon resolution. Nature. 2010; 464:1012–1017. [PubMed: 20393556]

27. Nutiu R, et al. Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument. Nat. Biotechnol. 2011; 29:659–664. [PubMed: 21706015]

28. Bentley DR, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008; 456:53–59. [PubMed: 18987734]

29. Carey J, Lowary PT, Uhlenbeck OC. Interaction of R17 coat protein with synthetic variants of its ribonucleic acid binding site. Biochemistry. 1983; 22:4723–4730. [PubMed: 6626527]

30. Lim F, David SP. Mutations that increase the affinity of a translational repressor for RNA. Nucleic Acids Res. 1994; 22:3748–3752. [PubMed: 7937087]

31. Valegård K, Murray JB, Stockley PG, Stonehouse NJ, Liljas L. Crystal structure of an RNA bacteriophage coat protein-operator complex. Nature. 1994; 371:623–626. [PubMed: 7523953]

32. Romaniuk PJ, Lowary P, Wu HN, Stormo G, Uhlenbeck OC. RNA binding site of R17 coat protein. Biochemistry. 1987; 26:1563–1568. [PubMed: 3297131]

33. Grahn E, et al. Structural basis of pyrimidine specificity in the MS2 RNA hairpin-coatprotein complex. RNA. 2001; 7:1616–1627. [PubMed: 11720290]

34. Bardwell VJ, Wickens M. Purification of RNA and RNA-protein complexes by an R17 coat protein affinity method. Nucleic Acids Res. 1990; 18:6587–6594. [PubMed: 1701242]

35. Bertrand E, et al. Localization of ASH1 mRNA particles in living yeast. 1998; 2:437–445.

36. Kivioja T, et al. Counting absolute numbers of molecules using unique molecular identifiers. Nat Meth. 2011; 9:72–74.

37. Greenleaf WJ, Frieda KL, Foster DA, Woodside MT, Block SM. Direct observation of hierarchical folding in single riboswitch aptamers. Science. 2008; 319:630–633. [PubMed: 18174398]

38. Hobson D, Uhlenbeck OC. Alanine Scanning of MS2 Coat Protein Reveals Protein– Phosphate Contacts Involved in Thermodynamic Hot Spots. Journal of Molecular Biology. 2006; 356:613–624. [PubMed: 16380130]

39. Varani G, McClain WH. The G·U wobble base pair. EMBO reports. 2000; 1:18–23. [PubMed: 11256617]

40. Valegârd K, et al. The three-dimensional structures of two complexes between recombinant MS2 capsids and RNA operator fragments reveal sequence-specific protein- RNA interactions. Journal of Molecular Biology. 1997; 270:724–738. [PubMed: 9245600]

41. Breen MS, Kemena C, Vlasov PK, Notredame C, Kondrashov FA. Epistasis as the primary factor in molecular evolution. Nature. 2012; 490:535–538. [PubMed: 23064225]

42. McCandlish DM, Rajon E, Shah P, Ding Y, Plotkin JB. The role of epistasis in protein evolution. Nature. 2013; 497:E1–E2. [PubMed: 23719465]

43. Weinreich DM. Darwinian Evolution Can Follow Only Very Few Mutational Paths toFitter Proteins. Science. 2006; 312:111–114. [PubMed: 16601193]

44. Bridgham JT, Ortlund EA, Thornton JW. An epistatic ratchet constrains the direction of glucocorticoid receptor evolution. Nature. 2009; 461:515–519. [PubMed: 19779450]

45. Natarajan C, et al. Epistasis Among Adaptive Mutations in Deer Mouse Hemoglobin. Science. 2013; 340:1324–1327. [PubMed: 23766324]

46. Rousset F, Pélandakis M, Solignac M. Evolution of compensatory substitutions through G.U intermediate state in Drosophila rRNA. Proceedings of the National Academy of Sciences of the United States of America. 1991; 88:10032–10036. [PubMed: 1946420]

47. Gell C, et al. Single-Molecule Fluorescence Resonance Energy Transfer Assays Reveal Heterogeneous Folding Ensembles in a Simple RNA Stem–Loop. Journal of Molecular Biology. 2008; 384:264–278. [PubMed: 18805425]

48. Licatalosi DD, et al. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. Nature. 2008; 456:464–469. [PubMed: 18978773]

49. Zhao J, et al. Genome-wide Identification of Polycomb-Associated RNAs by RIP-seq. Molecular Cell. 2010; 40:939–953. [PubMed: 21172659]

50. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNAbinding proteins and nucleosome position. Nat Meth. 2013; 10:1213–1218.

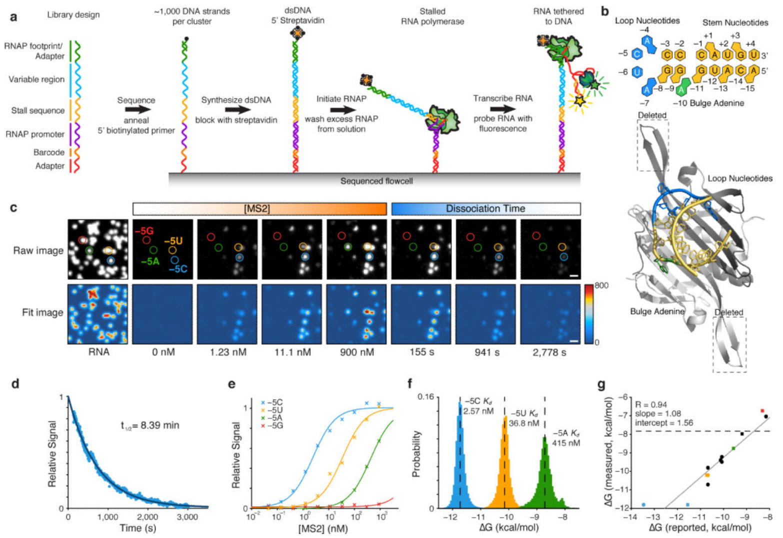

Figure 1. A massively parallel RNA array for quantitative, high-throughput biochemistry

(a) Steps for generating RNA tethered to DNA clusters on a high-throughput DNA sequencing flow cell. **(b)** Structure of the MS2 coat protein homodimer bound to the 19 nt hairpin RNA (PDB ID: 2BU1)[33]. **(c)** Images of fluorescently labeled MS2 bound to RNA clusters at increasing concentrations of protein and at time points following perfusion of unlabeled MS2 competitor. Below, fitted sum of Gaussians used to assign fluorescence to clusters. Scale bars (white) represent 2.5 μm. **(d)** Fluorescence decay of MS2 dissociating from clusters containing the consensus sequence (-5C) ($t_{1/2}$=8.39 minutes). **(e)** Fit binding curves to clusters labeled in panel (**c**). **(f)** The probability distribution of binding energies from all clusters with labeled variants; mean K_d = 2.57 nM, 36.8 nM, and 415 nM for the -5C, -5U, and -5A variants, respectively. **(g)** Correlation between binding energies reported in the literature and measured on the RNA array (squares, Carey et al.[29], circles, Romaniuk et al.[32]). (Dashed line indicates our affinity measurement cutoff.)

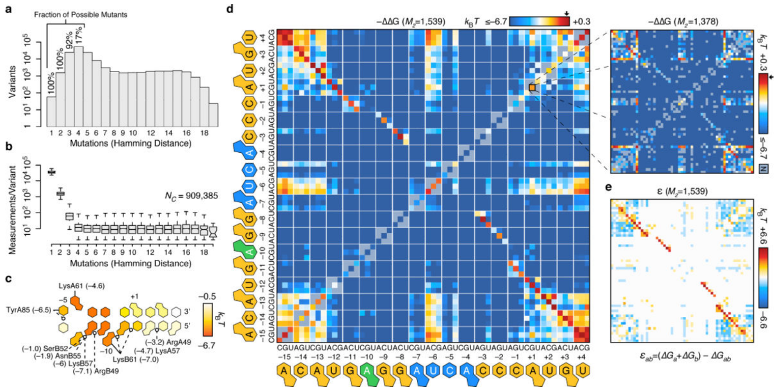

Figure 2. A quantitative map of MS2 binding across RNA sequence variants

(a) Distribution of observed RNA variants by number of mutations. **(b)** Clusters measured per molecular variant as a function of mutation number. A median of ~11 clusters are observed for sequences with ≥4 mutations. Affinities for the consensus sequence come from N_C=909,385 clusters. **(c)** Average −ΔΔ*G* of point mutations per position. The −ΔΔ*G* of alanine[38] substitutions to the MS2 binding surface are shown in parentheses (*k*B*T*). Solid and dashed lines represent base and phosphate interactions, respectively. **(d)** Matrix of −ΔΔ*G* for single and double mutants of the consensus sequence. Inset contains the matrix of −ΔΔ*G* for single and double mutants of the +1G variant. All energies are calculated relative to the consensus (-5C) sequence (arrow, −ΔΔ*G*=0), and the number of quality-filtered double mutants in each matrix is indicated (M_2). **(e)** Epistasis matrix derived from **(d)** allows *de novo* reconstruction of the hairpin structure.

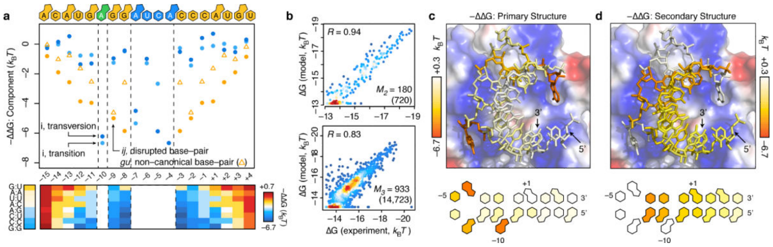

Figure 3. Binding affinity is dependent on primary sequence and secondary RNA structure

(a) Fit parameters for linear regression model showing position-specific contributions. Energetic components for all possible base pair combinations are shown below. **(b)** Predicted binding energies of variants with second (M_2) and third mutations (M_3) in both single- and double-stranded regions. Primary (i.e. mean energetic contributions of transitions and transversions) **(c)** and secondary **(d)** structure contributions to affinity derived from **a**, were mapped onto the hairpin (PDB ID: 1ZDH)[40].

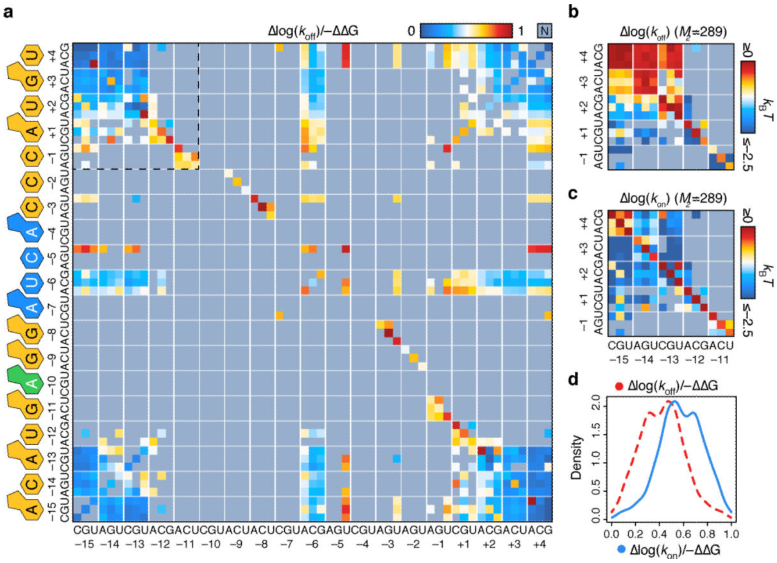

Figure 4. Sequence-specific contributions of association and dissociation rates to binding affinity

(a) Fractional contribution of dissociation rates for 31 single and 289 double mutants with measurable affinities and dissociation rates. Positions at the base of the hairpin are highlighted. **(b)** Δlog(k_{off}) and **(c)** Δlog(k_{on}) at the base of the hairpin. M_2 = number of qualityfiltered double mutants. **(d)** Distribution of fractional contributions of association (blue, μ=0.57) and dissociation (red, μ=0.43) rates to −ΔΔG for all measured mutants (N=3,029).

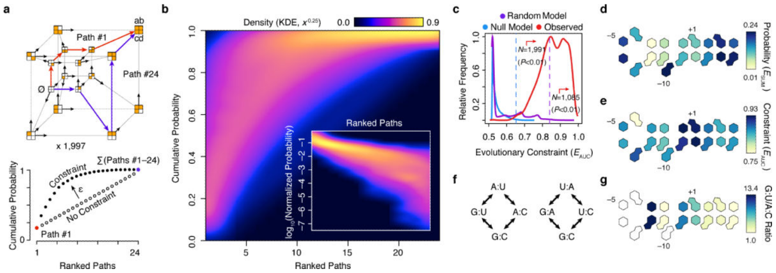

Figure 5. Evolutionary landscapes are highly constrained by biophysical requirements

(a) Tesseracts describe traversal probabilities for the complete set (*N*=24) of mutational paths between low and high-affinity variants within 4 mutations. The AUC of the cumulative probability of ranked paths measures evolutionary constraint (E_{AUC}), as modulated by epistasis (ε). **(b)** Density of cumulative probabilities for the ranked paths of 1,997 measured tesseracts. The fraction of the total path probabilities captured per individual path is shown as a function of path rank in the inset. The cumulative sum of these individual values is integrated to calculate E_{AUC}. **(c)** Distribution of E_{AUC} scores from observed tesseracts (red), tesseracts with uniform path probabilities (blue) and tesseracts with random affinities (purple) imply a highlystructured epistatic landscape. The number of variants significantly constrained ($P < 0.01$, Benjamini-Hochberg) is indicated for both models. Average evolutionary probability **(d)** and constraint **(e)** for paths with changes at each position of the hairpin. **(f)** Intermediate trajectories for base pair A:U→G:C and U:A→G:C transitions. **(g)** Probability ratio of evolutionary paths passing through G:U vs. A:C intermediates by base derived from 696 tesseracts with A:U→G:C base pair transformations.