nature biotechnology

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

Jason D Buenrostro[1,2,4], Carlos L Araya[1,4], Lauren M Chircus[1,3], Curtis J Layton[1], Howard Y Chang[2], Michael P Snyder[1] & William J Greenleaf[1]

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 a fluorescently labeled protein to >10^7 RNA targets generated on a flow cell surface by *in situ* transcription and intermolecular tethering of RNA to DNA. Studying the MS2 coat protein, we decompose the binding energy contributions from primary and secondary RNA structure, and observe that differences in affinity are often driven by sequence-specific changes in both association and dissociation 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 (RNA-MaP) provides generalizable insight into the biophysical basis and evolutionary consequences of sequence-function relationships.

RNA-protein interactions influence gene expression[1], viral assembly[2] and a wide variety of other critical biological processes. 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 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 (dsDNA), RNA substrates demonstrate diverse intramolecular interactions—including, mismatched base bulges, stem loops, pseudo knots, g-quartets, divalent cation interactions and noncanonical 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], which recently have been used to generate a catalog of RNA binding motifs[20]. Although powerful, selection and sequencing methods bias results toward 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 exist 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 total internal reflection fluorescence imaging for massively parallel DNA sequencing[28], to create a platform for direct, ultra-high-throughput measurement of RNA-protein interactions.

[1]Department of Genetics, Stanford University School of Medicine, Stanford, California, USA. [2]Program in Epithelial Biology and the Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, California, USA. [3]Department of Chemical and Systems Biology, Stanford University School of Medicine, Stanford, California, USA. [4]These authors contributed equally to this work. Correspondence should be addressed to W.J.G. (wjg@stanford.edu).

Received 21 January; accepted 19 March; published online 13 April 2014; doi:10.1038/nbt.2880

In addition, we have developed quantitative image analysis tools for large-scale analysis of these data, and measure both equilibrium binding constants and dissociation kinetics. 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 data sets enabling predictive models for affinity tuning, decomposition of binding energies between primary and secondary structures, and quantitative analysis of evolutionary trajectories across sequence space. We apply these methods to the coat protein of MS2, a bacteriophage that often infects *Escherichia coli*[2,29–33]. The MS2 coat protein and RNA hairpin have widespread applications in affinity purification[34], RNA imaging[35] and synthetic biology[10,11].

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** and **Supplementary Table 1**). To ensure multiple measurements of each RNA variant and reduce sequencing errors[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 ~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 present at a median of 15 locations 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 two-dimensional (2D) array.

After sequencing, we removed the sequenced DNA strand and regenerated dsDNA using DNA polymerase to extend a biotinylated primer. We then saturated the flow cell with streptavidin to create a

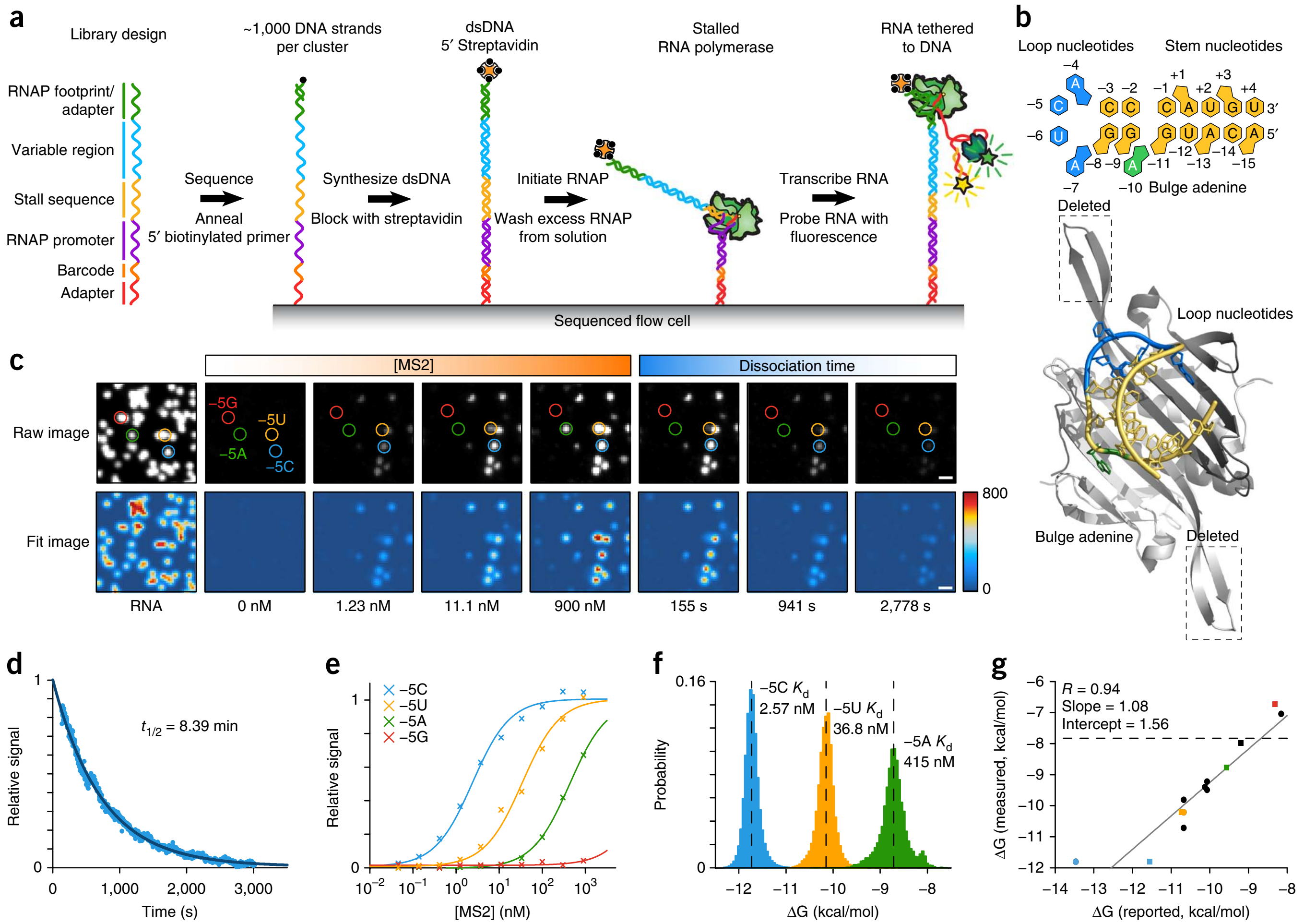

Figure 1 A massively parallel RNA array for quantitative, high-throughput biochemistry. (**a**) Steps for generating RNA tethered to DNA clusters on a high-throughput, DNA sequencing flow cell. (**b**) Structure of the MS2 coat protein homodimer bound to the 19-nt hairpin RNA (PDB ID: 2BU1)[33]. (**c**) 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, 2.5 µm. (**d**) Fluorescence decay of MS2 dissociating from clusters containing the consensus (–5C) sequence ($t_{1/2}$ = 8.39 min). (**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.)

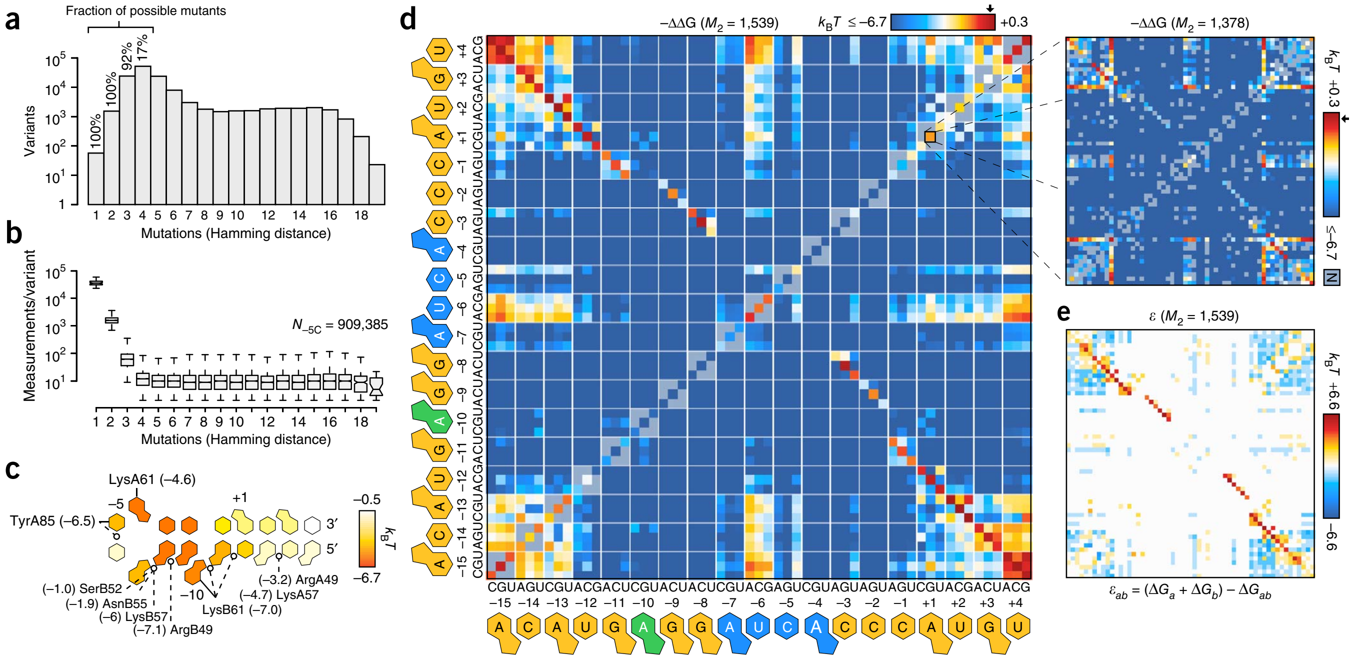

Figure 2 A quantitative map of MS2 binding across RNA sequence variants. (**a**) Distribution of observed RNA variants by number of mutations. (**b**) Clusters measured per molecular variant as a function of mutation number. A median of ~11 clusters are observed for sequences with ≥4 mutations. Affinities for the consensus (–5C) sequence come from N_{-5C} = 909,385 clusters. Box plots show median and upper/lower quartiles; whiskers show minimum and maximum. (**c**) Average $-\Delta\Delta G$ of point mutations per position. The $-\Delta\Delta G$ of alanine[38] substitutions to the MS2 binding surface are shown in parentheses (k_BT). Solid and dashed lines represent base and phosphate interactions, respectively. (**d**) Matrix of $-\Delta\Delta G$ for single and double mutants of the consensus sequence. Inset contains the matrix of $-\Delta\Delta G$ for single and double mutants of the +1G variant. All energies are calculated relative to the consensus sequence (arrow, $-\Delta\Delta G_{-5C} = 0$), and the number of quality-filtered double mutants in each matrix is indicated (M_2). Gray: no data (N). (**e**) Epistasis matrix derived from **d** allows *de novo* reconstruction of the hairpin structure.

terminal biotin-streptavidin roadblock on the 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* RNAP holoenzyme 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 by RNA polymerase (**Fig. 1a** and **Supplementary Fig. 3**). The resulting RNA array contained 1.2×10^7 distinct clonal RNA populations 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 SNAP-Surface 549 over the RNA array, and imaged bound MS2 protein at equilibrium using total internal reflection fluorescence at ten increasing concentrations. After the final measurement, we perfused 1.8 μM unlabeled MS2 protein and recorded the fluorescence decay caused by dissociation (**Fig. 1c** and **Supplementary Movie 1**). The high concentration of unlabeled MS2 protein blocks other binding sites on the array, preventing rebinding 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 Figs. 4 and 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** and **Supplementary Fig. 6**) and binding curves (**Fig. 1e,f** and **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 Tables 2** and **3**; for error estimation and quality control, **Supplementary Figs. 9** and **10**). To investigate how sequence variation in the RNA hairpin affects MS2 binding, we examined differential binding energies for all single mutants compared to the consensus sequence ($-\Delta\Delta G_{-5C} = 0\ k_BT$). 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

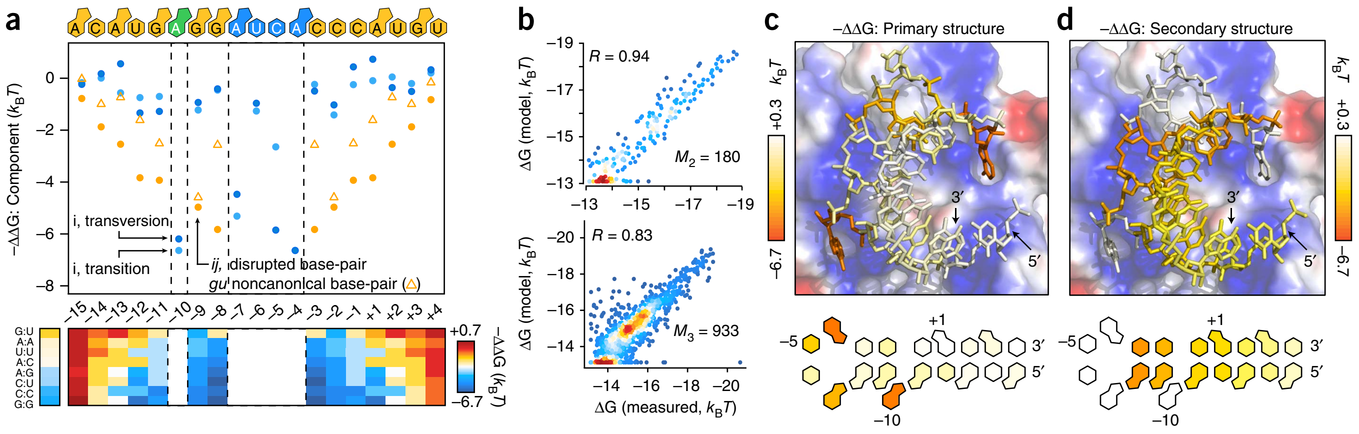

Figure 3 Decomposition of primary and secondary RNA structure determinants of binding affinity. (**a**) Fit parameters for linear regression model showing position-specific contributions. Energetic components for all possible noncanonical base-pair combinations are shown below. (**b**) Predicted binding energies of variants with second (M_2) and third mutations (M_3) in both single- and double-stranded regions. *R*, Pearson's correlation coefficient. (**c**,**d**) 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]. MS2 protein surface color represents electrostatic charge.

negative epistasis in noncompensating 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 noncanonical base-pairing interactions. These parameters were optimized jointly, to identify (by 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 noncanonical bases in isolation from primary sequence perturbations). To quantify the sensitivity for noncanonical base-pairing at positions in the hairpin stem, we trained the model eight separate times (once for each possible noncanonical pairing) with one free parameter representing the energetic cost of the respective noncanonical pairing. This refitting analysis allowed the model to incorporate a different energetic penalty for having noncanonical 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 noncanonical 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 noncanonical 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, whereas noncanonical 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 −Δ*ΔG* values for all mutants with measurable kinetic data. We calculated the energetic contributions to −Δ*ΔG* from changes in dissociation rates $[-\log(k_{\text{off}}^{\text{mutant}}/k_{\text{off}}^{\text{consensus}}) \overset{\text{def}}{=} \Delta\log(k_{\text{off}}^{\text{mutant}})]$, and inferred the contribution from changes in association rates, $[\log(k_{\text{on}}^{\text{mutant}}/k_{\text{on}}^{\text{consensus}}) \overset{\text{def}}{=} \Delta\log(k_{\text{on}}^{\text{mutant}})]$. 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 −Δ*ΔG* across single and double mutants (**Fig. 4a**). At the base of the hairpin, only a small fraction of −Δ*ΔG* 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 (**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 whereas 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 −Δ*ΔG* 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.

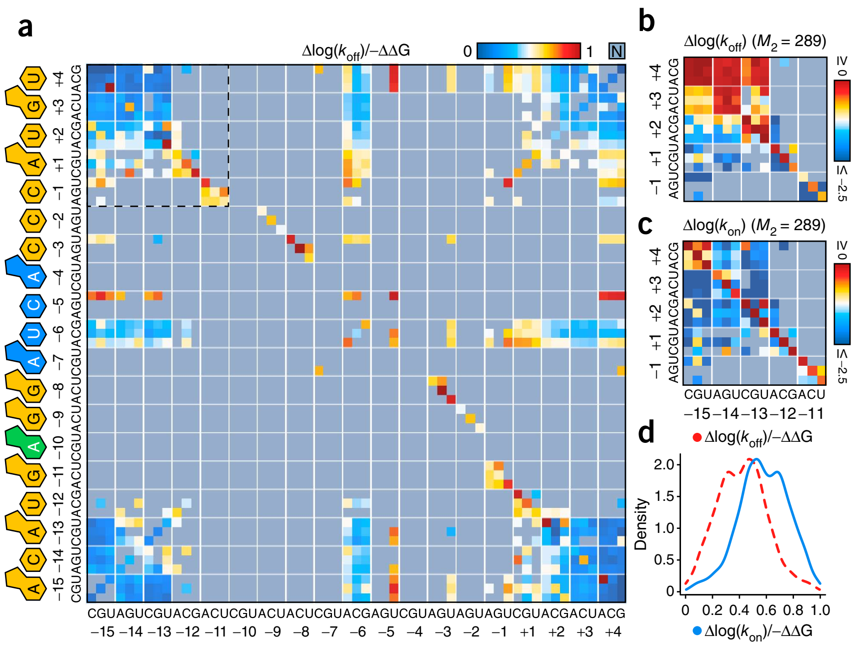

Figure 4 Sequence-specific contributions of association and dissociation rates to binding affinity. (**a**) Fractional contribution of dissociation rates for 31 single and 289 double mutants with measurable affinities and dissociation rates. Positions at the base of the hairpin are highlighted. Gray: no data (N). (**b**,**c**) $\Delta\log(k_{off})$ (**b**) and $\Delta\log(k_{on})$ (**c**) at the base of the hairpin. M_2 = number of quality-filtered double mutants. (**d**) Distribution of fractional contributions of association (blue, $\mu = 0.57$) and dissociation (red, $\mu = 0.43$) rates to $-\Delta\Delta G$ for all measured mutants ($N = 3{,}029$).

Quantitative analysis of evolutionary landscapes

We sought to understand how biophysical properties shape RNA sequence evolution toward higher binding affinity by examining the prevalence of epistasis, or differential mutational path probabilities caused by nonadditive 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 allowed us to examine how molecular evolution toward 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, and the 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 toward 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 nonepistatic landscape accounting for measurement errors (null model), or from a model that assumes a random distribution of affinities (**Fig. 5c**). These results indicate that distributions of affinity

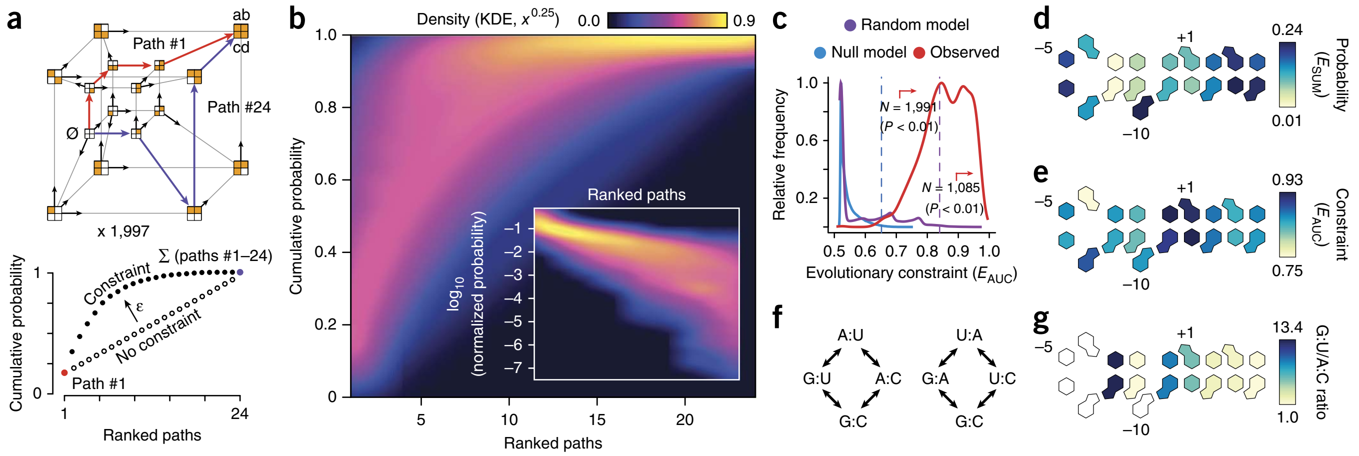

Figure 5 Evolutionary landscapes are highly constrained by biophysical requirements. (**a**) Tesseracts describe traversal probabilities for the complete set ($N = 24$) of mutational paths between low- and high-affinity variants within four mutations. The AUC of the cumulative probability of ranked paths 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}. KDE, kernel density estimation. (**c**) Distribution of E_{AUC} scores from observed tesseracts (red), tesseracts with uniform path probabilities (blue) and tesseracts with random affinities (purple) imply a highly structured epistatic landscape. The number of variants significantly constrained ($P < 0.01$, Benjamini-Hochberg) is indicated for both models. (**d**,**e**) 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 versus A:C intermediates by base derived from 696 tesseracts with A:U→G:C base pair transformations.

effects in mutational paths are highly structured, 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 nonuniform 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**; for data see **Supplementary Table 4**), implying that biophysical properties impose strong, systematic, structure-dependent effects on evolutionary trajectories.

By modeling evolutionary sequence preference of ribosomal RNA, Rousset *et al.*[46] observed that trajectories transitioning from A:U to G:C base pairs preferentially traverse G:U versus A:C intermediates and hypothesized this noncanonical base-pairing as a general mechanism for maintaining RNA-protein contacts in evolution. 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 underlying 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 Fig. 13**), we observed 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 intermolecular 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 on a large 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 data set, 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 quantification of the effect of primary sequence, secondary structure and noncanonical 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. Although 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 observed[46]. 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 useful 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 RNA generated *in vitro*, 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 through fluorescence resonance energy transfer (FRET). In addition, the sequencing platform is capable of generating DNA clusters >1 kb[50], enabling transcription of long RNAs and allowing investigations of long, noncoding 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.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. SRA: SRX495154.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

This work was supported by National Institutes of Health (NIH) NIH R01-HG004361 (to H.Y.C.); H.Y.C. is an Early Career Scientist of the Howard Hughes Medical Institute. J.D.B. and L.M.C. acknowledge support from the National Science Foundation Graduate Research Fellowships. J.D.B. also acknowledges support from NIH training grant T32HG000044. L.M.C. acknowledges support from NIH training grant T32GM067586. J.D.B. and L.M.C. each acknowledge support of a National Science Foundation graduate research fellowship. M.P.S. and C.L.A. acknowledge the NIH and the National Human Genome Research Institute (NHGRI) for funding through 5U54HG00455805. 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 quantification of cluster fluorescence. We also thank O.D. Phanstiel, M. Sikora and O. Cornejo for discussions on the modeling and evolutionary analyses.

AUTHOR CONTRIBUTIONS

W.J.G., J.D.B. and C.J.L. conceived of the method. J.D.B. developed the RNA display protocol. J.D.B. and L.M.C. designed and performed on-chip assays. L.M.C. designed and performed the protein purification and *in vitro* 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.

1. Keene, J.D. RNA regulons: coordination of post-transcriptional events. *Nat. Rev. Genet.* **8**, 533–543 (2007).
2. Carey, J., Cameron, V., De Haseth, P.L. & Uhlenbeck, O.C. Sequence-specific interaction of R17 coat protein with its ribonucleic acid binding site. *Biochemistry* **22**, 2601–2610 (1983).
3. Tsvetanova, N.G., Klass, D.M., Salzman, J. & Brown, P.O. Proteome-wide search reveals unexpected RNA-binding proteins in *Saccharomyces cerevisiae*. *PLoS ONE* **5**, e12671 (2010).
4. Scherrer, T., Mittal, N., Janga, S.C. & Gerber, A.P. A screen for RNA-binding proteins in yeast indicates dual functions for many enzymes. *PLoS ONE* **5**, e15499 (2010).
5. Butter, F., Scheibe, M., Morl, M. & Mann, M. Unbiased RNA-protein interaction screen by quantitative proteomics. *Proc. Natl. Acad. Sci. USA* **106**, 10626–10631 (2009).
6. Castello, A. *et al.* Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell* **149**, 1393–1406 (2012).
7. Wang, K.C. *et al.* A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature* **472**, 120–124 (2011).
8. Tsai, M.C. *et al.* Long noncoding RNA as modular scaffold of histone modification complexes. *Science* **329**, 689–693 (2010).
9. Guttman, M. & Rinn, J.L. Modular regulatory principles of large non-coding RNAs. *Nature* **482**, 339–346 (2012).
10. Culler, S.J., Hoff, K.G. & Smolke, C.D. Reprogramming cellular behavior with RNA controllers responsive to endogenous proteins. *Science* **330**, 1251–1255 (2010).
11. Ausländer, S., Ausländer, D., Müller, M., Wieland, M. & Fussenegger, M. Programmable single-cell mammalian biocomputers. *Nature* **487**, 123–127 (2012).
12. SantaLucia, J. & Turner, D.H. Measuring the thermodynamics of RNA secondary structure formation. *Biopolymers* **44**, 309–319 (1997).
13. Kertesz, M. *et al.* Genome-wide measurement of RNA secondary structure in yeast. *Nature* **467**, 103–107 (2010).
14. Ding, Y. *et al. In vivo* genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature* **505**, 696–700 (2014).
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* **505**, 701–705 (2014).
16. Ban, N., Nissen, P., Hansen, J., Moore, P.B. & Steitz, T.A. The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science* **289**, 905–920 (2000).
17. Wan, Y., Kertesz, M., Spitale, R.C., Segal, E. & Chang, H.Y. Understanding the transcriptome through RNA structure. *Nat. Rev. Genet.* **12**, 641–655 (2011).
18. Martin, L. *et al.* Systematic reconstruction of RNA functional motifs with high-throughput microfluidics. *Nat. Methods* **9**, 1192–1194 (2012).
19. Ray, D. *et al.* Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nat. Biotechnol.* **27**, 667–670 (2009).
20. Ray, D. *et al.* A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **499**, 172–177 (2013).
21. Araya, C.L. *et al.* A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. *Proc. Natl. Acad. Sci. USA* **109**, 16858–16863 (2012).
22. Pitt, J.N. & Ferre-D'Amare, A.R. Rapid construction of empirical RNA fitness landscapes. *Science* **330**, 376–379 (2010).
23. Guenther, U.-P. *et al.* Hidden specificity in an apparently nonspecific RNA-binding protein. *Nature* **502**, 385–388 (2013).
24. Matzas, M. *et al.* High-fidelity gene synthesis by retrieval of sequence-verified DNA identified using high-throughput pyrosequencing. *Nat. Biotechnol.* **28**, 1291–1294 (2010).
25. Myllykangas, S., Buenrostro, J.D., Natsoulis, G., Bell, J.M. & Ji, H.P. Efficient targeted resequencing of human germline and cancer genomes by oligonucleotide-selective sequencing. *Nat. Biotechnol.* **29**, 1024–1027 (2011).
26. Uemura, S. *et al.* Real-time tRNA transit on single translating ribosomes at codon resolution. *Nature* **464**, 1012–1017 (2010).
27. Nutiu, R. *et al.* Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument. *Nat. Biotechnol.* **29**, 659–664 (2011).
28. Bentley, D.R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
29. Carey, J., Lowary, P.T. & Uhlenbeck, O.C. Interaction of R17 coat protein with synthetic variants of its ribonucleic acid binding site. *Biochemistry* **22**, 4723–4730 (1983).
30. Lim, F. & David, S.P. Mutations that increase the affinity of a translational repressor for RNA. *Nucleic Acids Res.* **22**, 3748–3752 (1994).
31. Valegård, K., Murray, J.B., Stockley, P.G., Stonehouse, N.J. & Liljas, L. Crystal structure of an RNA bacteriophage coat protein-operator complex. *Nature* **371**, 623–626 (1994).
32. Romaniuk, P.J., Lowary, P., Wu, H.N., Stormo, G. & Uhlenbeck, O.C. RNA binding site of R17 coat protein. *Biochemistry* **26**, 1563–1568 (1987).
33. Grahn, E. *et al.* Structural basis of pyrimidine specificity in the MS2 RNA hairpin-coat-protein complex. *RNA* **7**, 1616–1627 (2001).
34. Bardwell, V.J. & Wickens, M. Purification of RNA and RNA-protein complexes by an R17 coat protein affinity method. *Nucleic Acids Res.* **18**, 6587–6594 (1990).
35. Bertrand, E. *et al.* Localization of ASH1 mRNA particles in living yeast. *Mol. Cell* **2**, 437–445 (1998).
36. Kivioja, T. *et al.* Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods* **9**, 72–74 (2012).
37. Greenleaf, W.J., Frieda, K.L., Foster, D.A., Woodside, M.T. & Block, S.M. Direct observation of hierarchical folding in single riboswitch aptamers. *Science* **319**, 630–633 (2008).
38. Hobson, D. & Uhlenbeck, O.C. Alanine scanning of MS2 coat protein reveals protein–phosphate contacts involved in thermodynamic hot spots. *J. Mol. Biol.* **356**, 613–624 (2006).
39. Varani, G. & McClain, W.H. The G·U wobble base pair. *EMBO Rep.* **1**, 18–23 (2000).
40. Valegård, K. *et al.* The three-dimensional structures of two complexes between recombinant MS2 capsids and RNA operator fragments reveal sequence-specific protein-RNA interactions. *J. Mol. Biol.* **270**, 724–738 (1997).
41. Breen, M.S., Kemena, C., Vlasov, P.K., Notredame, C. & Kondrashov, F.A. Epistasis as the primary factor in molecular evolution. *Nature* **490**, 535–538 (2012).
42. McCandlish, D.M., Rajon, E., Shah, P., Ding, Y. & Plotkin, J.B. The role of epistasis in protein evolution. *Nature* **497**, E1–E2 (2013).
43. Weinreich, D.M. Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science* **312**, 111–114 (2006).
44. Bridgham, J.T., Ortlund, E.A. & Thornton, J.W. An epistatic ratchet constrains the direction of glucocorticoid receptor evolution. *Nature* **461**, 515–519 (2009).
45. Natarajan, C. *et al.* Epistasis among adaptive mutations in deer mouse hemoglobin. *Science* **340**, 1324–1327 (2013).
46. Rousset, F., Pélandakis, M. & Solignac, M. Evolution of compensatory substitutions through G.U intermediate state in *Drosophila* rRNA. *Proc. Natl. Acad. Sci. USA* **88**, 10032–10036 (1991).
47. Gell, C. *et al.* Single-molecule fluorescence resonance energy transfer assays reveal heterogeneous folding ensembles in a simple RNA stem–loop. *J. Mol. Biol.* **384**, 264–278 (2008).
48. Licatalosi, D.D. *et al.* HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature* **456**, 464–469 (2008).
49. Zhao, J. *et al.* Genome-wide identification of polycomb-associated RNAs by RIP-seq. *Mol. Cell* **40**, 939–953 (2010).
50. Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y. & Greenleaf, W.J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).

ONLINE METHODS

Library design and construction. To generate a high-density RNA array, we designed a custom DNA library containing a barcode, *E. coli* RNAP promoter, RNAP stall sequence, constant region and degenerate MS2 hairpin sequence (**Supplementary Fig. 1**). The reverse complement 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 nonconsensus 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. The 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 1× NEBnext PCR Mix, 1.25 µM oligos C and D, and 0.6× 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 two 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 used to transform BL21(DE3) cells (NEB), and starter cultures of transformed cells were grown overnight in a rotator at 37 °C in 10 ml Luria-Bertani medium (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 h. SNAPtag-MS2 expression was induced with 0.5 mM IPTG for 5 h at 22 °C, and then cells were collected by centrifugation at 4,000 r.p.m. for 15 min 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 1 h, the flow-through was reapplied to the column five times, and the column was washed two 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 by 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 min in 50 mM Tris pH 8.0, 100 mM NaCl, 0.1% Tween 20, 1 mM DTT and 10 µM SNAP-Surface 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 1× 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 8,100 nM in TMKG buffer (TMK buffer, 10% glycerol, 100 µg/ml BSA) for 1.75 h 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 h. 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 ($signal_{\mathrm{nitrocellulose}}$) (which binds protein and therefore MS2-RNA complexes) over the total signal on both the nitrocellulose and the positively charged nylon ($signal_{\mathrm{+Nylon}}$) (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 min 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 single-stranded DNA, 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 (1× 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 each ATP, GTP and UTP, 0.015 mg/ml RNAP (Sigma-saturated holoenzyme from Epicentre) and 0.01% Tween-20) for 30 min 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 1× T7A1 reaction buffer plus 2.5 µM each ATP, GTP and UTP. Finally, 1 mM NTPs (ATP, CTP,

doi:10.1038/nbt.2880

GTP and UTP) in 1× T7A1 reaction buffer were added for 30 min at 37 °C to allow transcription to proceed. After transcription, RNAP remained stalled at the 5′ biotin-streptavidin 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 Figs. 1** and **3**). Based on cluster fluorescence intensities, we observed an RNA synthesis efficiency of ~30–40%. We also annealed an unlabeled MS2_3′ block oligo to the constant region between the hairpin and the RNAP footprint to help 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 3× increasing concentrations starting at 0.046 nM and ending at 900 nM for a total of ten binding images. For each measurement, we waited 1 h 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.

Image processing. Cluster positions, including tile, *x* position and *y* position were extracted from the FASTQ sequencing data. Cluster positions were then cross-correlated with acquired images to define a global *x*/*y* offset (**Supplementary Figs. 4** and **5**). Before quantification, saturated pixels within the image were masked. After cross-correlation, images were broken into smaller sub images (24 × 24 pixels) and fit to a sum of overlapping 2D Gaussians (defined by the sequencing cluster centers) using linear least squares (**Supplementary Figs. 4** and **5**):

$$F_b + \sum_{k=0}^{n} A_k \exp\left(-\left(\frac{(x - x_0 - x_{off,k} - x_{global})^2}{2\sigma_k^2} + \frac{(y - y_0 - y_{off,k} - y_{global})^2}{2\sigma_k^2}\right)\right)$$

This fit was repeated for all images (120 tiles per GAIIx sequencing lane), which included all 120 RNA images (Alexa Fluor 647) and 26 × 120 binding and dissociation images (SNAP-Surface 549). Following single cluster fits to 2D Gaussians, we grouped data by cluster ID and calculated the fit fluorescence using the fit parameters A_k= amplitude and σ_k= standard deviation with the following equation:

$$F_{int} = 2\pi A_k \sigma_k^2$$

Barcode handling and sequence alignment. A random barcode and bottlenecking approach was used to reduce sequencing and measurement error by ensuring replicate clusters. We first removed barcodes with homopolymers of 15 consecutive bases or barcodes that were represented three times or less; 495,822 of clusters were removed (4.29%). To reduce sequencing errors, we grouped sequences by identical barcodes. Hairpin variants that were represented twice or less among a set of identical barcodes were assigned as sequencing errors and removed from further analysis; 990,053 clusters were removed (8.58%). Using these stringent cutoffs, 10,059,446 clusters remained. Remaining clusters were aligned to the MS2 consensus sequence using a Smith-Waterman algorithm.

Fitting K_d and k_{off}. Data were normalized to the total RNA per cluster, quantified using the Alexa Fluor 647 labeled DNA oligo. Outlier data were estimated using median absolute deviation (MAD) and removed. Integrated fluorescence values from the each cluster were fit to a binding curve with the equation:

$$F_{obs} = \frac{F_{max}}{1 + \frac{K_d}{x}} + F_{min}$$

where F_{obs} = observed fluorescence, F_{max} = fit maximum fluorescence, F_{min} = fit minimum fluorescence, K_d = affinity constant and x = concentration of MS2. Outliers were estimated using MAD and removed. Reported binding energy values are the median Gibbs free energy (kcal/mol). Measurement error was estimated using bootstrapped 95% confidence intervals on this median.

For dissociation calculations, we found that the consensus sequence (–5C variant) had 9% residual fluorescence (standard deviation of 4%) after ~90 min of dissociation, which likely attributed to nonspecific fluorescence. Therefore, the maximum and minimum intensities of each cluster were normalized by bound MS2 at 900 nM and remaining fluorescence after ~90 min of dissociation. Clusters of identical sequence were merged using image time stamps and median fluorescence values per variant at each time point. Each variant and their associated merged values were used to fit dissociation constants using the following equation:

$$F_{obs} = (1 - F_{ns})\exp(-k_{off}\,x) + F_{ns}$$

where F_{obs} = observed fluorescence, F_{ns} = nonspecific background fluorescence, k_{off} = dissociation rate and x = dissociation time.

Differential affinity calculations and quality filtering. For each unique sequence (N = 148,184), we computed the $-\Delta\Delta G$ of binding relative to the consensus sequence, and quality-filtered our data for analysis by removing sequences for which the computed range of the 95% confidence interval for $-\Delta\Delta G$ was greater than 1 k_BT. In addition, we retained sequences if the upper bound of the 95% confidence interval lies below $-\Delta G < 13.12236\ k_BT$ (K_d > 2,000 nM), thresholding their binding affinity to K_d = 2,000 nM. As a result, retained sequences either bind the MS2 coat protein at K_d < 2,000 nM, with affinity estimates within 0.5 k_BT (on average) of the upper/lower bound affinities, or bind at $K_d \geq$ 2,000 nM. To determine fit quality, we calculated mean square error (MSE) of each single cluster fit and determined the median MSE for each variant and removed variants (N = 194) with an MSE > 0.025 (**Supplementary Fig. 7**). Finally, we removed a set of variants (N = 184) with $-\Delta\Delta G > 0.9\ k_BT$, comprising a population of clusters with nonconverging fits, yielding a final set of quality-approved variants of N = 129,248. Binding energies and quality metrics for approved variants are provided in **Supplementary Table 2**. Similarly, we quality-filtered our kinetics data approving 3,029 variants with vetted binding energies and qualities (above), K_d < 500 nM and half-lives ($t_{1/2}$) > 0.5 min. Measured dissociation and inferred association rates are provided in **Supplementary Table 3**.

Analysis of intramolecular epistasis. We calculated intramolecular epistasis scores (ε_{ab}) for the complete set of double mutants following a simple additive model of neutrality:

$$\varepsilon_{ab} = (\Delta G_a + \Delta G_b) - \Delta G_{ab}$$

Whereby ΔG_{ab}, ΔG_a and ΔG_b represent the measured Gibbs free energy of binding MS2 for the variant harboring mutations *a* and *b*, the variant with the point mutation *a*, and the variant with the *b* point mutation, respectively. As such, positive ε_{ab} values reflect sequences in which the composite mutations (*a*, *b*) yield higher binding affinity than would be expected from their individual effects.

To derive the secondary structure of the MS2 RNA hairpin solely from affinity measurements, we calculated the mean epistasis score per *i,j* position pair (**Fig. 2e** and **Supplementary Fig. 12a**). Selecting pairs of positions with mean epistasis scores ≥1 s.d. from zero and with the reciprocal, highest mean epistasis scores accurately identified the structure of base pairs in the RNA hairpin.

Modeling affinity effects of mutation. To infer the energetic contribution of primary and secondary structure defects conferred by mutations at each position of the MS2 hairpin, we first classified positions as base-paired or single-stranded on the basis of epistasis signatures (described above). For each position *i*, we generated a binary (*1,0*) matrix detailing the presence or absence of specific primary and secondary structure defects (columns) from variants (rows) harboring mutations exclusively at relevant *i* or *i,j* positions, for single-stranded or base-paired positions, respectively (**Supplementary Fig. 12b**). Specifically, we incorporated annotation terms for (i) transitions and (ii) transversions at *i*, for (iii) transitions and (iv) transversions at *j*, as well as for (v) an individual noncanonical base-pair at *i,j*, and (vi) base-pair disruption as defined by a mismatched base pair (not canonical, and not noncanonical) at *i,j*. With the exception of annotation terms for transitions and transversions at *i*, all terms were set to zero for positions in single-stranded regions. In addition, for each position, an entry for the consensus (unmutated) sequence was included in which all annotation terms were set to zero. As such, the complete matrix (for all positions) included data from 121 input variants with ≤2 mutations ($M_0 = 1$, $M_1 = 57$, $M_2 = 63$).

For each position *i*, we modeled the $-\Delta\Delta G$ of binding (relative to the consensus) of the annotated variants with a linear regression on the annotation terms to derive the energetic contribution of each primary and secondary structure defect (per position):

$$-\Delta\Delta G_i = \omega_i^1 \cdot i_{transition} + \omega_i^2 \cdot i_{transversion} + \omega_{i,j}^3 \cdot j_{transition} + \omega_{i,j}^4 \cdot j_{transversion} + \omega_{i,j}^5 \cdot i, j_{non-canonical} + \omega_{i,j}^6 \cdot i, j_{disrupt}$$

To predict the ΔG of binding of distant, untrained sequences, we selected double ($M_2 = 720$) and triple ($M_3 = 14{,}723$) mutants that combine mutations in base-paired and unpaired positions. This criterion was implemented to reduce cooperative effects from mutations in adjacent base pairs, a structural feature not captured by the model. We predicted the binding energy (ΔG) of the selected sequences through two approaches. In the first (*ab initio*) approach, we predicted the $\Delta\Delta G$ by compiling the predicted, conferred effect of the specific primary and secondary defects on the ΔG of the consensus RNA hairpin. In a second (*mutation impact*) approach, we estimated the $\Delta\Delta G$ resulting from primary and secondary structure defects conferred by additional mutations based on the ΔG of sequences harboring the complementary set of mutations. For example, to predict the ΔG_{ab} of binding of double mutant *ab*, we predicted the $\Delta\Delta G_{b,a}$ introduced by *b* on the ΔG_a of binding of *a*, and the $\Delta\Delta G_{a,b}$ introduced by *a* on ΔG_b. We then estimated the ΔG_{ab} as the mean of $\Delta G_a + \Delta\Delta G_{b,a}$ and $\Delta G_b + \Delta\Delta G_{a,b}$. For both approaches, the predicted ΔGs were floored for $K_d \geq 2{,}000$ nM.

We evaluated the accuracy of the predicted ΔG for double ($M_2 = 180$) and triple ($M_3 = 933$) mutants with $K_d < 2{,}000$ nM. The *ab initio* approach predicted the ΔG of doubles and triples with $R_2 = 0.91$ and $R_3 = 0.81$, whereas the mutation impact provides improved predictions at $R_2 = 0.94$ and $R_3 = 0.83$, respectively. Predictions from the mutation impact approach are shown in **Figure 3b**.

Analysis of evolutionary paths. To examine the probability of evolution along distinct mutational paths, we first generated a graph connecting individual hairpin sequences within one mutation of each other, yielding a mutation graph with 104,395 hairpin sequences (nodes) connected through 620,100 single-point mutations (edges). We scanned this graph for pairs of low- and high-affinity sequences separated by four mutations for which binding energies of all intermediates were measured (i.e., sequences containing all subsets of the four mutations separating the low- and high-affinity sequences). We required high-affinity variants with $-\Delta\Delta G \geq -1$ and low-affinity variants with $-\Delta\Delta G \geq -6.5$, thus allowing a broad range of affinity gains between low- and high-affinity variants, as well as a broad range of sequences.

The data from the mutation graph allowed us to reconstruct the complete sets of mutational paths for 1,997 pairs of low- and high-affinity sequences, allowing us to examine the relative probabilities of 47,928 mutational trajectories describing the serial acquisition of four mutations. Each complete set contains $N = 24$ (4!) mutational paths, which can be mathematically arranged as tesseracts (**Fig. 5b**). We defined the probability of mutating from variant *ø* to variant *a* within each tesseract ($M_{ø\rightarrow a}$) to be the binding energy of *a* divided by the sum of the binding energies of all sequences within one mutation from *ø* as well as the binding energy of *ø* (effectively, the probability of binding *a* versus *ø* and all sequences in the tesseract within one mutation of *ø*). Therefore, because there are four independent mutations (*a*, *b*, *c*, *d*) separating a low-affinity variant (*ø*) from a high-affinity variant (*abcd*), each of which can be acquired first, the probability of $M_{ø\rightarrow a}$ can be calculated as:

$$M_{\phi\rightarrow a} = \frac{e^{-\Delta\Delta G_a / k_B T}}{\sum_{i=\{\phi,a,b,c,d\}} e^{-\Delta\Delta G_i / k_B T}}$$

We define a mutational path as the serial acquisition of mutations in a specific order, and the serial acquisition of mutations in any order as a mutational trajectory. Therefore, the probability of traversing a mutational path ($P_{ø\rightarrow a\rightarrow ab\rightarrow abc\rightarrow abcd}$) is given by the product of the individual mutation probabilities (i.e., $M_{ø\rightarrow a}$, $M_{a\rightarrow ab}$, $M_{ab\rightarrow abc}$, $M_{abc\rightarrow abcd}$):

$$P_{\phi\rightarrow a\rightarrow ab\rightarrow abc\rightarrow abcd} = M_{\phi\rightarrow a} {}^{*} M_{a\rightarrow ab} {}^{*} M_{ab\rightarrow abc} {}^{*} M_{abc\rightarrow abcd}$$

and the probability of realization (E_{SUM}) of a mutational trajectory ($T_{ø\rightarrow abcd}$) is given by the sum of the probabilities of all mutational paths linking *ø* and *abcd*. Because our model incorporates the probability of not mutating at each node, the probabilities of mutational trajectories do not equal 1. Therefore, we normalized the mutational path probabilities by E_{SUM} for the evolutionary constraint analyses.

Following Weinreich *et al.*[43], we examined constraint in the accessibility of distinct evolutionary paths by modeling the cumulative (normalized) probability of ranked mutational paths. We extended these approaches by calculating the AUC of the cumulative probability curve to capture constraint as a single metric. This metric of evolutionary constraint (E_{AUC}) was tightly correlated with entropy in the mutational path probabilities within each tesseract (Spearman's $\rho = 0.995$). Evolutionary probability and constraint metrics are reported in **Supplementary Table 4**.

Null and random models of evolutionary constraint. The null model of evolutionary constraint is that the differences in mutational path probabilities arise solely from measurement error. We generated 100 bootstrapped null models for each tesseract by (i) randomly selecting a mutational path within the tesseract to obtain a reference ΔG for each Hamming distance from the starting (low-affinity) and ending (high-affinity) sequences and (ii) randomly selecting an affinity for each variant at each Hamming distance within the 95% confidence interval of the reference ΔG for that Hamming distance. We generated 100 null models for each observed tesseract ($N = 1{,}997$), resulting in 199,700 null models. We calculated the significance of observed evolutionary constraint (E_{AUC}) scores from the distribution of null model scores and corrected for multiple hypothesis testing (Benjamini-Hochberg). 1,991 (99.7%) of the observed tesseracts are significantly ($P < 0.01$) constrained compared to the null model (**Fig. 5c**).

The random model of evolutionary constraint is that the differences in mutational path probabilities arise solely from random distribution of mutational path probabilities (and therefore binding energies) within tesseracts. We generated 199,700 random models by arbitrarily assigning observed binding energies to variants within each model and calculating evolutionary constraint (E_{AUC}). As a comparison, we calculated the significance of observed evolutionary constraint (E_{AUC}) scores from the distribution of random model scores, correcting for multiple hypothesis testing (Benjamini-Hochberg). 1,085 (54.3%) of the observed tesseracts are significantly constrained at $P < 0.01$ in the random model (**Fig. 5c**).