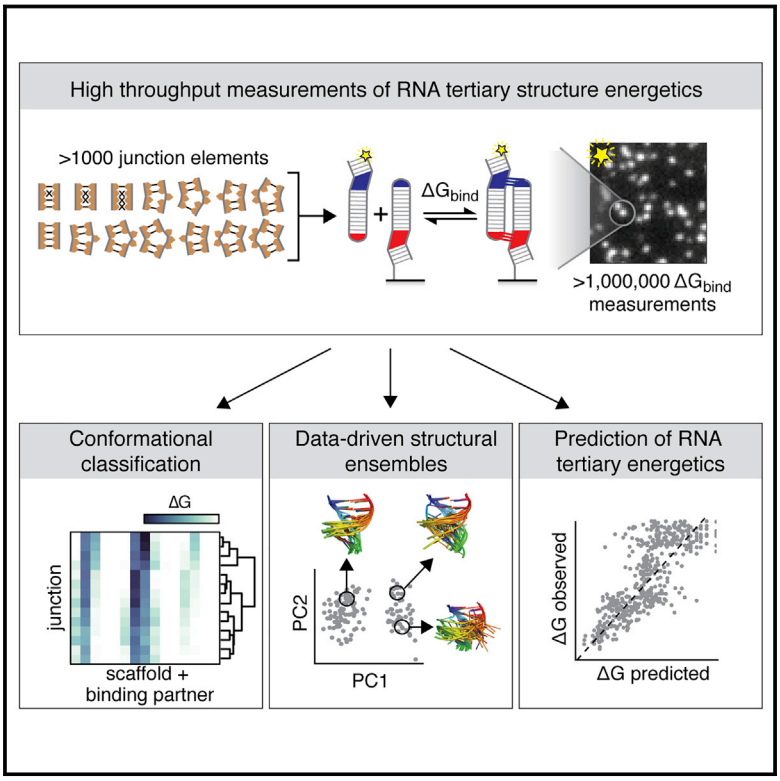
Article

Cell

High-Throughput Investigation of Diverse Junction Elements in RNA Tertiary Folding

Graphical Abstract



Highlights

- Characterization of >1,000 RNA junctions with thermodynamic fingerprints
- Arrangement and sequence of non-WC residues dictate junction conformations
- Dynamic ensembles deduced by integrating static structural and energetic data
- Effects on RNA tertiary folding energetics are predicted by structural ensembles

Authors

Sarah Knight Denny, Namita Bisaria, Joseph David Yesselman, Rhiju Das, Daniel Herschlag, William James Greenleaf

Correspondence

herschla@stanford.edu (D.H.), wjg@stanford.edu (W.J.G.)

In Brief

Characterizing the thermodynamic fingerprints of >1,000 RNA junctions reveals principles for how RNA sequence affects tertiary assembly energetics, highlighting a path toward tertiary folding prediction by integrating static structural and dynamic energetic information.

Denny et al., 2018, Cell *174*, 377–390
July 12, 2018
https://doi.org/10.1016/j.cell.2018.05.038

CellPress

Article

Cell

High-Throughput Investigation of Diverse Junction Elements in RNA Tertiary Folding

Sarah Knight Denny,[1,10] Namita Bisaria,[2,9,10] Joseph David Yesselman,[2] Rhiju Das,[2,3] Daniel Herschlag,[2,4,5,6,*] and William James Greenleaf[1,3,7,8,11,*]

[1]Program in Biophysics, Stanford University, Stanford, CA 94305, USA
[2]Department of Biochemistry, Stanford University School of Medicine, Stanford, CA 94305, USA
[3]Department of Applied Physics, Stanford University, Stanford, CA 94305, USA
[4]Department of Chemistry, Stanford University, Stanford, CA 94305, USA
[5]Department of Chemical Engineering, Stanford University, Stanford, CA 94305, USA
[6]ChEM-H Institute, Stanford University, Stanford, CA 94305, USA
[7]Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305, USA
[8]Chan Zuckerberg Biohub, San Francisco, CA 94158, USA
[9]Present address: Whitehead Institute for Biomedical Research, Cambridge, MA 02142, USA
[10]These authors contributed equally
[11]Lead Contact
*Correspondence: herschla@stanford.edu (D.H.), wjg@stanford.edu (W.J.G.)
https://doi.org/10.1016/j.cell.2018.05.038

SUMMARY

RNAs fold into defined tertiary structures to function in critical biological processes. While quantitative models can predict RNA secondary structure stability, we are still unable to predict the thermodynamic stability of RNA tertiary structure. Here, we probe conformational preferences of diverse RNA two-way junctions to develop a predictive model for the formation of RNA tertiary structure. We quantitatively measured tertiary assembly energetics of >1,000 of RNA junctions inserted in multiple structural scaffolds to generate a "thermodynamic fingerprint" for each junction. Thermodynamic fingerprints enabled comparison of junction conformational preferences, revealing principles for how sequence influences 3-dimensional conformations. Utilizing fingerprints of junctions with known crystal structures, we generated ensembles for related junctions that predicted their thermodynamic effects on assembly formation. This work reveals sequence-structure-energetic relationships in RNA, demonstrates the capacity for diverse compensation strategies within tertiary structures, and provides a path to quantitative modeling of RNA folding energetics based on "ensemble modularity."

INTRODUCTION

Structured RNAs are critical in diverse biological processes, including the regulation of gene expression, protein translation, and pre-mRNA splicing (Moore, 2005; Noller, 2005). To function in these biological processes, RNAs must fold into intricate, 3-dimensional (3D) structures. In this folding process, an RNA sequence typically folds hierarchically, compacting into secondary structure elements before the formation of tertiary conformations (Brion and Westhof, 1997; Tinoco and Bustamante, 1999; see also Chauhan and Woodson, 2008; Strulson et al., 2014). The thermodynamic stability of a secondary structure can be predicted by summing the free energies associated with individual secondary structure elements such as base pair (bp) steps (Turner et al., 1988). In contrast, it is not yet possible to quantitatively predict the stability of forming a tertiary structure from a given secondary structure despite the crucial importance of this final step in RNA folding and in biological function.

RNA tertiary structures are comprised of helices, junctions, and sparsely distributed tertiary contacts that form between distal interaction interfaces. Tertiary contacts form when these interfaces come into proximity with specific geometric constraints, and the likelihood of forming the contact depends on the intervening helices and junctions that position the interaction interfaces (Chu et al., 2009; Mustoe et al., 2014). Thus, predicting the free energy of forming any tertiary interaction requires a full accounting of its 3D context, making it a substantially more difficult problem than that of secondary structure formation.

Over the past decades, RNA researchers have developed the perspective that RNA's architecture is "modular" and that it might be possible to build arbitrary tertiary structures by assembling the 3D structures of each of its constituent elements, analogous to assembling LEGOs. For example, certain tetraloop/tetraloop-receptor (TL/TLR) tertiary contacts are structurally superimposable across distinct RNAs, a commonality that would be surprising if the tertiary contact did not inherently favor a single structural conformation (Wu et al., 2012). Additionally, some RNA junctions impose common conformations on their emanating helices even when isolated from their original RNA context, further supporting structural modularity (Klein et al., 2001; Murchie et al., 1998; Thomson and Lilley, 1999). On the other hand, structured RNAs can undergo dramatic conformational transitions, facilitated by the inherent flexibility of helices and junctions that explore a range of conformational states. Understanding these

spatial preferences is integral to quantitatively determining the likelihood of forming a tertiary interaction—and, ultimately, to building a predictive model for RNA-tertiary-structure formation from the properties of individual elements.

Current techniques to characterize the conformational preferences of RNA elements are either unable to capture their dynamic behavior or are low-throughput (Salmon et al., 2014; Shi et al., 2014). Thus, our current understanding of RNA ensemble behavior is limited to a small number of RNA junction elements—too few to extrapolate to more general rules for how primary sequence and secondary structure may enforce particular 3D behaviors. This bottleneck is even more acute due to the combinatorial complexity of RNA, with thousands of uncharacterized secondary structure motifs in diverse biological RNAs.

To address these issues, we developed an approach that couples the conformational ensemble of an RNA element to a quantitative, thermodynamic readout using a platform that enables massively parallel measurements (Buenrostro et al., 2014). The readout is the binding equilibrium between two structured RNAs that form a tertiary assembly (tectoRNA) through the interaction of two TL/TLR tertiary contacts (Figure 1A) (Jaeger and Leontis, 2000). The likelihood of forming the tectoRNA assembly reflects the inherent conformational properties of its constituent RNA elements—i.e., the helices, junctions, and tertiary contacts that compose the assembly—as is true in the formation of any tertiary structure. In a separate study, we used this system to understand the conformational preferences of RNA helices with distinct Watson Crick (WC) bp compositions (Yesselman et al., 2018) and found that small differences in the conformational ensembles of bp step elements could transduce into substantial changes in the thermodynamic stability of the tectoRNA (> 2 kcal/mol). These results support the sensitivity of the tectoRNA system to changes in the alignment of the tertiary contacts, likely enabled by the system's small size and the rigidity of the tertiary contacts.

Here, we apply this platform to understand the conformational behavior of RNA junctions, which serve as the flexible "pivot points" responsible for a large fraction of the dynamic behavior of an RNA. We focus on two-way junctions such as bulges and internal loops, which comprise about 70% of biological junctions (Cruz and Westhof 2009; Petrov et al., 2013). Each junction was integrated into multiple assemblies that enforce different conformational requirements for productive tertiary assembly formation. The set of binding measurements for each junction forms a multidimensional "thermodynamic fingerprint" that is influenced by the junction's 3D behavior, allowing broad comparison of conformational preferences of two-way junction variants. This analysis revealed that junction conformations are dictated by the number and arrangement of unpaired residues, with the sequence identity of non-WC-paired residues having an additional role. Further, we were able to infer dynamic ensembles for previously uncharacterized junctions by integrating these thermodynamic fingerprints with existing crystallographic structural data. This ensemble description of junctions was combined with our prior ensemble model for WC bp steps to predict the tertiary folding energetics of these model RNA assemblies.

RESULTS

High-Throughput Measurement of Tertiary Assembly Formation with TectoRNAs

We chose the tectoRNA system (Figure 1A) to dissect energetic contributions to RNA-tertiary-structure formation (Jaeger and Leontis, 2000). A tectoRNA assembly consists of two structured RNAs that form a heterodimer through two discrete TL/TLR tertiary interactions: the ubiquitous GAAA-11nt TL/TLR and the engineered GGAA-R1 TL/TLR (Figures 1A and S1A) (Costa and Michel, 1995; Geary et al., 2008). The likelihood of forming both tertiary contacts depends on the alignment of the tertiary interaction interfaces, thus enabling quantitative assessment of tertiary structure formation through the binding affinity of the heterodimer (Figure 1B).

To test the effect of RNA two-way junctions on the formation of this tectoRNA assembly, we inserted designed junction sequences into one piece of the tectoRNA heterodimer, the "chip piece," which was immobilized on a surface. The other binding partner, the "flow piece," was free in solution and contained a fluorophore used to measure the extent of complex formation (Figure 1C). Each junction was inserted into a set of "chip scaffolds" that varied in sequence and in the number of bps between the two tertiary contacts (Figures 1D and S1B). By using multiple chip scaffolds and three distinct flow pieces, we could measure the effect of each inserted element on formation of a diverse set of assemblies.

We designed junctions that comprehensively sampled symmetric and asymmetric loops up to three in length (N = 1,328; Figure S1B). In addition, we included a set of junction sequences characterized by X-ray crystallography (N = 359) to relate energetic measurements in the tectoRNA system to structural behavior. Junctions were categorized into secondary structure classes according to nomenclature developed in (Bailor et al., 2010), indicated by "NxM," where N and M are the number of non-WC-paired residues on the 5′ and 3′ side of the hairpin loop, respectively (Figures 1D and S1B). Inserted elements replaced the number of possible paired residues within the junction: possible pairs are either WC or noncanonical pairs and do not include bulged residues.

To measure tectoRNA assembly at high throughput, the designed chip piece variants were synthesized as DNA, sequenced, and *in situ* transcribed, resulting in RNA that remains tethered to its sequence-identified DNA template on the surface of a sequencing flow cell (Figure S1D; STAR Methods; She et al., 2017). Increasing concentrations of the fluorescently labeled flow piece were introduced to the flow cell, and the fluorescence of the bound flow piece at each cluster was measured after waiting sufficient time for equilibration (Figure 1E). This set of fluorescent values was fit to a binding isotherm for each cluster of RNA to obtain the dissociation constant (K_d) and the free energy of binding ($\Delta G = RT \log[K_d]$) of the tectoRNA flow piece to each chip piece variant (Figure 1F).

Measurements of K_d values using different chips were highly reproducible ($R^2 = 0.96$; Figure 1G). K_d values spanned a range of several orders of magnitude (K_d values of 1–5,000 nM; ΔG values of −12.0 to −7.1 kcal/mol). Error estimates within an experiment were less than 1.4-fold (0.2 kcal/mol) for the

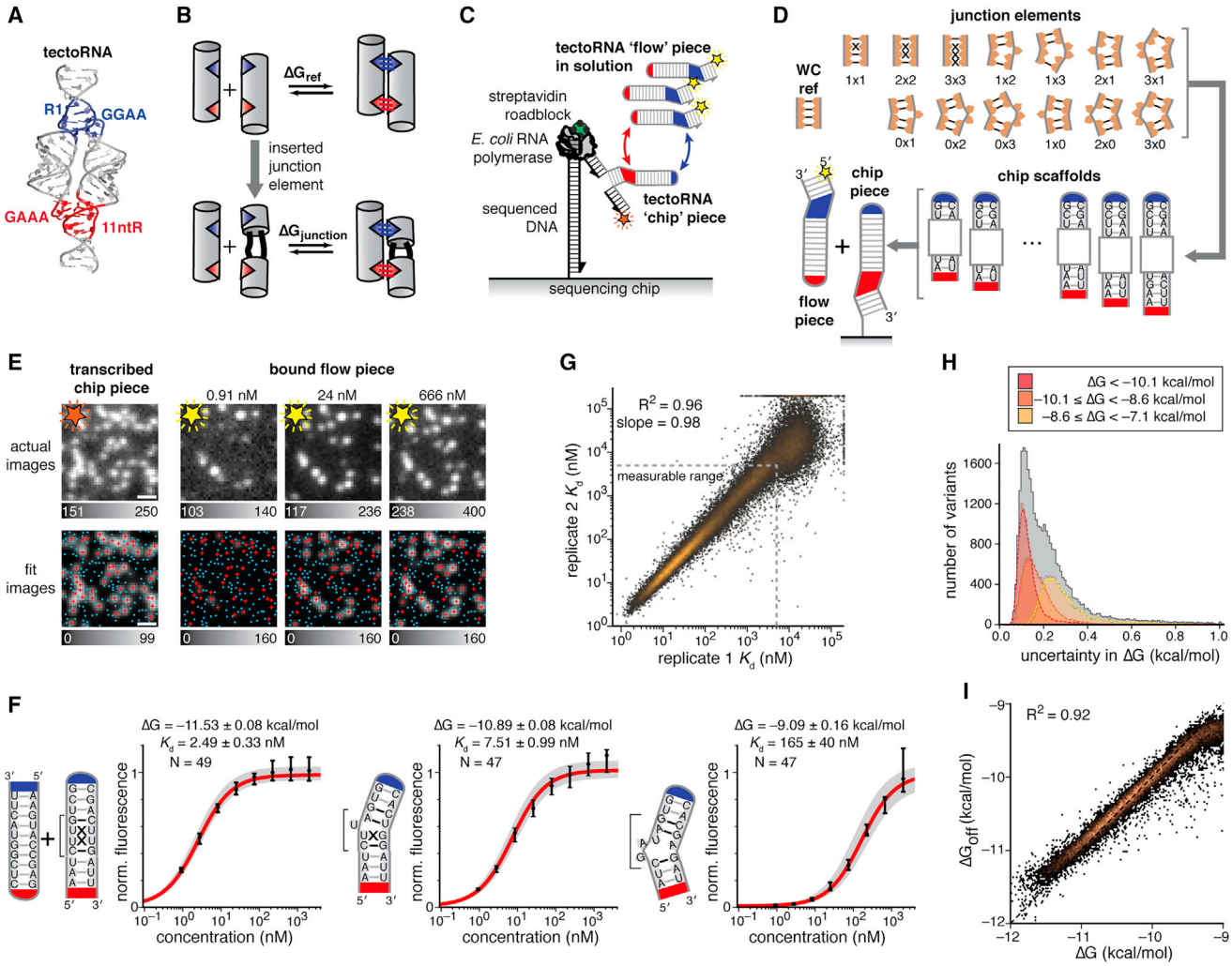


Figure 1. High-Throughput Characterization of RNA Junctions Using TectoRNAs

(A) TectoRNA homodimer structure (PDB: 2ADT) with two tetraloop/tetraloop receptors (TL/TLRs). The tectoRNA heterodimer used in this study replaces one of these TL/TLRs with the GGAA-R1 TL/TLR (blue) (Geary et al., 2008), while the other is the same as in the homodimer version, the GAAA-11nt TL/TLR (red).

(B) Schematic of tectoRNA complex formation with and without an inserted junction element.

(C) Schematic of experimental setup with the *in situ* transcribed "chip" piece and the "flow" piece in solution.

(D) Schematic of the tectoRNA library design.

(E) Observed (top) and fit (bottom) images of fluorescent RNA clusters immobilized on the sequencing chip surface. Fluorescent-labeled oligo hybridized to clusters of *in situ*-transcribed RNA (left). Binding of fluorescent-labeled flow piece to chip-piece clusters at three concentrations of the flow piece binding series (right). Known cluster centers are indicated (bottom); red crosses and cyan dots show clusters with and without an RNAP initiation site, respectively. Scale bar, 2.5 μm.

(F) Representative binding curves for three chip piece variants to a common 10 bp flow piece (far left). Free energy of binding (ΔG), dissociation constant (K_d), and number of clusters measured (N) are indicated for each variant. Error bars represent bootstrapped 95% CIs on quantified fluorescence from all single clusters associated with each molecular variant; gray area indicates the 95% CI on the fit parameters.

(G) Scatterplot of the binding affinity measurements of tectoRNA library measured in two replicate experiments. Measurable range corresponds to K_d values ≤ 5,000 nM.

(H) Distribution of estimated error on the fit ΔG values (95% CI). Gray denotes all variants with measurable affinity; colors denote bins based on the value of ΔG.

(I) Scatterplot of ΔG_{off} versus measured binding affinity (ΔG) across tectoRNA variants. $\Delta G_{off} = RT \log(k_{off}/c)$, where k_{off} is the measured dissociation rate constant and c is the fixed association rate constant ($c = 5.8 \times 10^4$ $M^{-1}s^{-1}$); see also Figures S4B and S4C). Dotted black line indicates $\Delta G = \Delta G_{off}$.

See also Figures S1 and S3.

large majority of stable binders (95% CI; Figures 1H), with values becoming more precise with increasing number of measurements and with stronger affinity (Figures 1H, S2A, and S2B). The ability to measure K_d values over a large dynamic range was enabled by analysis methods that used saturation fluorescence values for strong binders in estimating binding affinity for low-affinity variants (STAR Methods; Figures S2C–S2H).

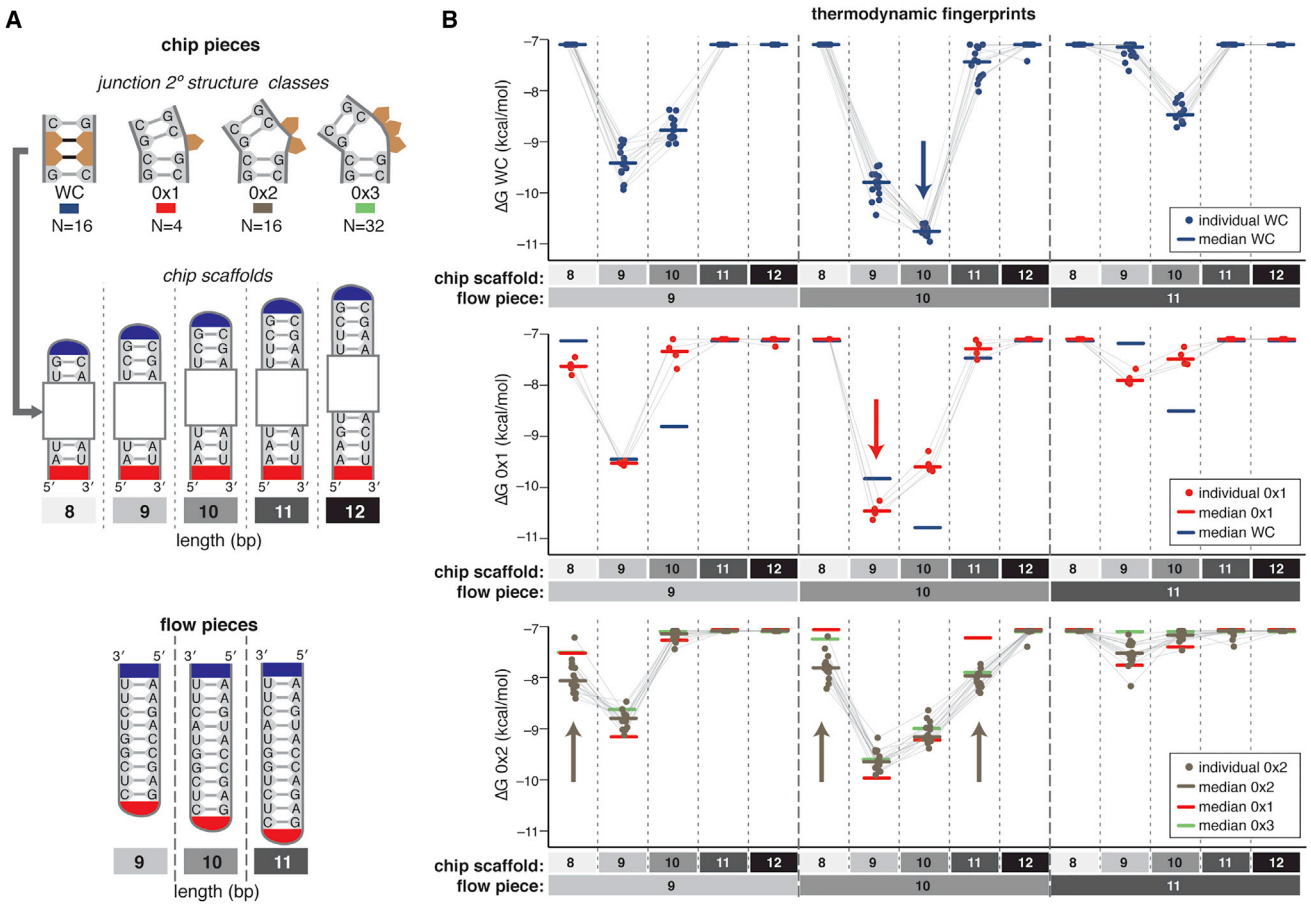


Figure 2. Thermodynamic Fingerprints of Junction Elements through Variation of Flow- and Chip-Piece Helix Length

(A) Junction elements are inserted into the indicated chip scaffolds to form chip pieces (top) and measured with three flow pieces (bottom).

(B) Thermodynamic fingerprints of inserted elements. Individual junction sequences (points) and the median value across junction sequences (horizontal lines) are shown for each flow/chip context. Blue and red arrows indicate the context with the lowest affinity for WC pairs (blue) or 0x1 motifs (red), respectively. Tan arrows (bottom) indicate contexts in which having a 0x2 motif is stabilizing compared to 0x1 or WC motifs. For 0x3 junctions, only the median value is shown for simplicity. See also Figure S4.

We assessed our measured affinities under different Mg^{2+} conditions and found consistent ΔΔG values between tectoRNA variants at 5 mM and 30 mM Mg^{2+} (Figure S3A). Subsequent assays were carried out at the higher 30 mM Mg^{2+} concentration as it gave a larger dynamic range to observe larger destabilizing effects.

In addition to binding affinity, we measured the dissociation rate constants, which showed a strong correlation and linear relationship to the thermodynamic measurements (R^2 = 0.92; Figures 1I and S3B). This observation supports a kinetic model in which the two tertiary contacts of the tectoRNA form sequentially, with the association of the first contact being rate limiting and independent of the structural context imposed by the diverse junctions and helical elements composing the tectoRNA chip pieces (Figures S3B and S3C).

TectoRNA Association Depends on the Alignment of Both Tertiary Contact Interfaces

To dissect how the tertiary contacts contribute to the tertiary stability of tectoRNAs, we included chip piece variants with mutated tetraloops (GAAA to GGAA, N = 6,198 scaffolds) and substantially altered tetraloop receptors (11ntR to tandem GC bps, N = 29 scaffolds). All these variants were significantly destabilized relative to the wild-type tertiary contacts, with very few of these mutated variants having apparent binding in the measurable range (ΔG > –7.1 for 99.5% of variants; Figures S3D–S3G). These controls demonstrate that both tertiary contacts are required for tectoRNA assembly formation within our concentration range.

We changed the relative positions of the tertiary contacts, while maintaining tertiary contacts themselves, by varying the length of the helical segment of the flow and chip pieces (Figure 2B, top; N = 16 sequence variants per length). The original lengths of 10 bp for the flow and chip pieces gave the highest affinity, and shortening or lengthening the chip piece by two bps was sufficient to ablate measurable binding (Figure 2B, top, middle). The tectoRNA affinity depended on both the flow piece and chip piece lengths (Figure 2B, top), with each combination presumably giving a different fraction of states that allow productive assembly formation.

We next tested whether there was interdependence between the energetic contribution of the tertiary contact interactions and the conformational effect of changing their alignment. The type of tertiary contact used in the tectoRNA system is known to form with specific geometric constraints (tetraloop-tetraloop receptors; [Wu et al., 2012]), leading to the simple expectation that each contact forms the same interactions across contexts (Bisaria et al., 2017). By this model, a point mutation in the tertiary contact would have the same effect across different RNA contexts. Conversely, mutating a residue in a flexible contact with multiple sets of possible interactions would have different effects in contexts that favor different substates of the tertiary contact. To distinguish these models, we made point mutants in the tertiary contact receptor in the tectoRNA chip piece (11ntR to 11nt-A4U and 11nt-U9G). We found a constant energetic effect on binding of these mutations across >250 tectoRNA scaffolds with varying length helices and with inserted junctions (Figures S4A and S4B), supporting a common set of tertiary interactions across these tectoRNA variants independent of the helix and junction elements between them. Consequently, the thermodynamic effects of inserted junctions should predominantly arise from alterations in the alignment of the tertiary contacts, and not from differences in the energetic contribution of the 11ntR contact itself.

Thermodynamic Fingerprints Reveal Distinct Behaviors of Two-Way Junctions

Inserting a junction element into the tectoRNA will affect the thermodynamic stability of the assembly if the junction's underlying conformational ensemble alters the population of the states that can simultaneously form both tertiary contacts. Because the relationship between conformational preferences and thermodynamic effects is not exact—e.g., two junctions could have the same thermodynamic effect even if they have different conformational preferences—we inserted each junction into multiple tectoRNA scaffolds. Each scaffold applies a different conformational constraint on the formation of the assembly, and we reasoned that measuring junction effects across multiple scaffolds would increase our power to distinguish conformational behavior. We refer to this multidimensional set of effects as the junction's "thermodynamic fingerprint."

To illustrate the utility of thermodynamic fingerprint analysis, we compared the effects of inserting junctions with bulged residues (0x1 motifs; N = 4) to WC pairs (N = 16) across scaffolds of different lengths (Figures 2A and 2B). Introduction of a bulged residue was destabilizing by ~1 kcal/mol in the 10 bp flow/10 bp chip context (Figure 2B), while it was stabilizing in a shorter 9 bp chip scaffold relative to the WC pairs (Figure 2B). Presumably, the bulge changes the conformational ensemble of the chip piece, resulting in misalignment of the two tertiary contacts in the most stable WC context (10/10 bp) but better alignment of the tertiary contacts in the alternate 10/9 bp context.

The thermodynamic fingerprints of 0x2 and 0x3 bulges followed the pattern of the 0x1 motif but had increased stability in both the short (8 bp) and long (11 bp) chip scaffolds compared to the 0x1 and WC junctions (Figure 2B). These differences suggest a broader range of conformations explored by these larger bulge motifs than WC bps and 0x1 bulges, allowing complex formation in these otherwise misaligned scaffolds. Further supporting this increased flexibility, the 0x2 and 0x3 bulges were more weakly bound than the WC pairs and 0x1 bulges in the most stable 10/9 and 10/10 bp contexts (Figure 2B).

To ensure that the relationships between fingerprints are not specific to the tertiary interaction used in the tectoRNA assembly, we repeated a subset of the thermodynamic fingerprints with a substantially altered tertiary contact receptor (C7.2) in place of the wild-type receptor 11ntR (Costa and Michel, 1997) (Figure S4C). The clustering of the thermodynamic fingerprints was largely reproduced between the 11ntR and C7.2 receptor, with the majority (59%) of junction fingerprints that clustered together in the 11ntR context also clustering together in the altered receptor context (Figure S4D). This value approached the 66% of junctions coclustering if the 11ntR fingerprints were reclustered after adding values sampled from measurement error. These observations are consistent with generality of the thermodynamic fingerprints beyond the specific tertiary assembly.

Arrangement of Unpaired Residues in Junctions Enforces Conformational Preferences

To understand the role of secondary structure in defining the conformational preferences of two-way junctions, the affinity fingerprints of 12 secondary structure classes (i.e., MxN categories) were compared (Figure 3A). Each profile is the average of all junction sequences within a secondary structure class. Hierarchical clustering of these profiles grouped the secondary structure classes by the similarity of their affinity fingerprints (Figure 3A), revealing that junctions with the same number of unopposed residues had similar average affinity profiles. The fingerprints of WC, 1x1, 2x2, and 3x3 junctions cluster together and are distinct from the 0x1, 1x0, 1x2, and 2x1 junctions, demonstrating distinct behavior for junctions with no unopposed residues and those with one unopposed residue (Figure 3A). These data support the model from Al-Hashimi and colleagues that junctions with shared topological constraints have similar conformational preferences (Bailor et al., 2010).

Analysis of individual thermodynamic fingerprints revealed a significant additional role of primary sequence in determining conformational behavior. Within each secondary structure class, we found substantial deviations in affinity among different sequence variants, as is shown in Figure 3B, with spread up to ~4 kcal/mol for different 3x3 junction sequences (Figures 3B and 3C; STAR Methods). The magnitude of sequence-specific deviations increased as the number of non-WC-opposing residues increased (Figure 3C), indicating that the identity of mismatched residues can have a large effect on the three-dimensional conformational behaviors of junctions. Overall, junction behavior is driven by the number and arrangement of unpaired residues, as well as the sequence identity of non-WC paired residues.

Purine-Pyrimidine Content of Mismatches Underlies Differences in Conformational Behavior

We next focused on the sequence determinants of the conformational behavior of single mismatch motifs (1x1) by obtaining thermodynamic fingerprints for each of the 12 possible

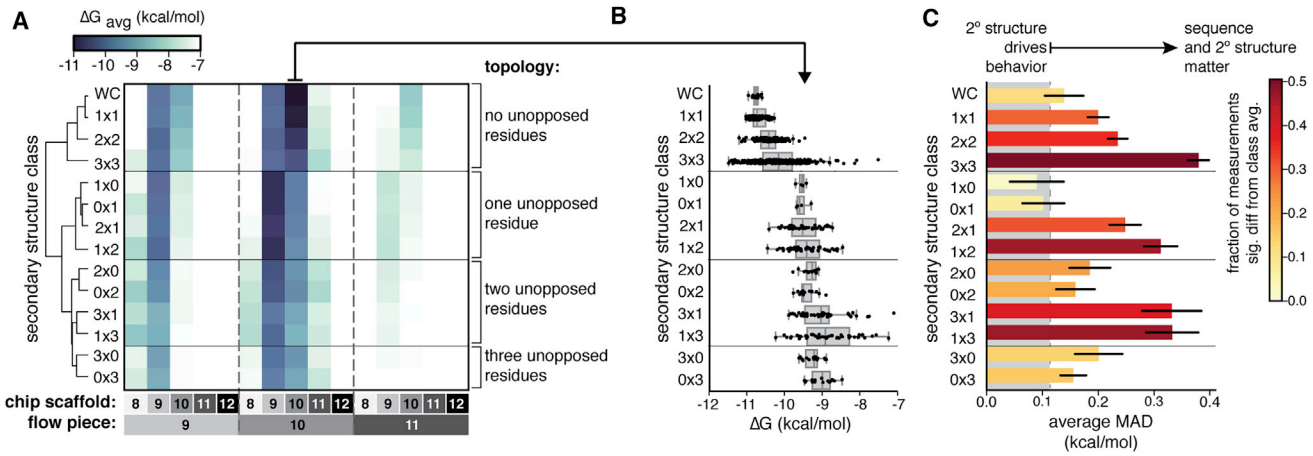


Figure 3. Topology and Sequence Drive Conformational Behavior of Junctions

(A) Heatmap of the hierarchically clustered, average thermodynamic fingerprints of each secondary structure class.

(B) Affinity measurements of individual junctions in the 10/10 bp flow/chip context.

(C) Deviation of individual junction sequences from the average profile of their secondary structure class. The MAD was calculated between each sequence and its class average profile. Average MAD of junction sequences within each class is shown (error bars are bootstrapped 95% CI). Color of the bars indicates the fraction of measurements that are significantly different than the average profile.

mismatches, each in two positions, with the other position occupied by one of the four canonical WC bps (N = 96 total; Figure 4A); fully WC-base paired motifs were also included for comparison (N = 16).

Clustering of the thermodynamic fingerprints identified seven classes that accounted for variation among these motifs (k-means clustering; average mean absolute deviation [MAD] within clusters of 0.12 kcal/mol, Figure S5A; Figures 4B and 4C). This clustering analysis revealed a strong distinction between purine-purine and pyrimidine-pyrimidine mismatch thermodynamic fingerprints: clusters 1–3 contained predominantly GG, AA, and AG or GA mismatches, respectively, whereas the most distal clusters, 6 and 7, were highly enriched in CC and CU or UC mismatches, respectively (Figures 4C–4E). These distinct thermodynamic effects suggest different conformational behavior for these mismatch types relative to the conformational behavior of WC pairs. The WC and purine-pyrimidine mismatches were distributed across the central three clusters and exhibited more similar fingerprints overall, implying similar conformational behavior among these sequences (Figures 4C–4E). The clustering pattern was not associated with the effect of each mismatch on the secondary structure stability of a duplex (Figure S5B), demonstrating that these thermodynamic effects likely do not arise from partial unfolding of the secondary structure.

To understand the physical behavior responsible for the distinct thermodynamic profiles of mismatches, we noted that the effects were similar across scaffolds that changed the position of the insertion (i.e., chip scaffolds 2–4 or 5–7 in Figures 4B and 4C). When mismatches were integrated into even more positions in helices of length 8–11 bp, again, the effect of each mismatch deviated very little across positions (< 0.2 kcal/mol for >90% of measurements) in all of these helix length contexts (Figure S5C). This observation was especially striking for AA mismatches in the 11 bp chip scaffolds (10 bp flow piece), which stabilized the tectoRNA assembly by ~1.2 kcal/mol relative to WC pairs in all positions, suggesting that AA mismatches have access to conformational states that allow the helix to shorten to compensate for the misalignment in the 11 bp scaffolds. Position-independent effects are consistent with the mismatches inducing a structural perturbation along the helical axis, such as a rotation or an increase in total contour length of the phosphate backbone, resulting in the same relative alignment of the tertiary contacts regardless of the mismatch position. This type of structural perturbation can also explain the observation that mismatch motifs were 5′ to 3′ interchangeable (e.g., GA mismatches had similar thermodynamic effects as AG mismatches and, similarly, for UC and CU mismatches; Figure 4E).

Overall, our analysis reveals simplifying rules for single-mismatch behavior, with seven classes accounting for the conformational behavior of the 112 mismatched motifs and WC pairs. We hypothesized that more classes would be required to account for the behavior of 2x2 and 3x3 motifs, as there is more sequence-specific variation within these secondary structure classes (Figure 3C). For 2x2 motifs, k-means clustering revealed 20 conformational classes that describe the thermodynamic fingerprints of these 144 motifs, and this number increases by more than 2-fold for 3x3 motifs (Figures S5A and S5D). The thermodynamic fingerprints of 2x2 purine-purine and pyrimidine-pyrimidine double mismatches were most distinct from the WC average profile, just as they were most distinct for the 1x1 mismatches (Figure S5E). Nevertheless, 2x2 and 3x3 sequences also had new thermodynamic behavior from 1x1 mismatches (Figures S5E and S5F). For example, a subset of the 2x2 motif sequences has position-dependent effects, and position dependence was even more pronounced for 3x3 junctions (Figures S5G). These observations suggest that the larger internal loops exhibit more complex conformational behavior that includes directional bending of the helix.

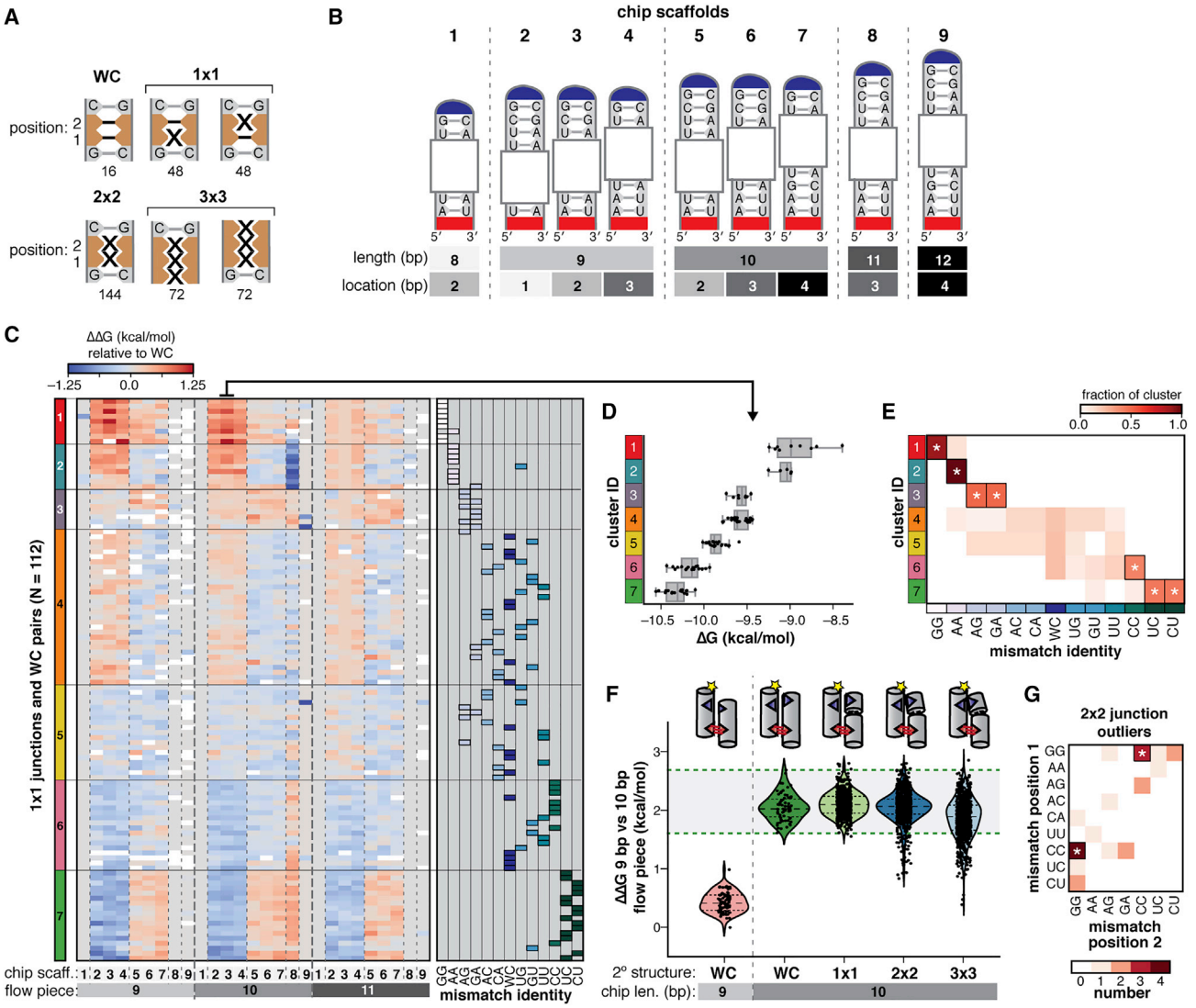


Figure 4. Identity of Mismatch Pairs Leads to Distinct Thermodynamic Behavior

(A) Schematic of mismatched junction elements. The number of junction sequences tested within each class is indicated.

(B) Chip scaffolds in which junctions were inserted. Scaffolds vary in length (8–12 bp) and location, indicated by the number of bps between the receptor and the junction element.

(C) Heatmap of the clustered thermodynamic fingerprints of 112 individual 1x1 junctions and WC elements, across the nine scaffolds defined in (B), and measured with three different flow pieces. Affinity is shown relative to the average profile of WC motifs across the same flow/chip contexts. White indicates missing measurement. Clusters are indicated by the left color bar. Heatmap on right indicates the mismatch present in each of the individual junctions, or "WC" if no mismatch present. "Chip scaff." is chip scaffold, as in (B).

(D) Individual affinity measurements of junctions inserted in a single structural context.

(E) The fractional representation of each mismatch type within each cluster. Outlined boxes with white asterisks indicate significant enrichment above expected fraction by chance (adjusted p value < 0.05; see STAR Methods).

(F) Points and violin plots show the difference in affinity between the 9 and 10 bp flow piece for a set of chip pieces. WC pairs and mismatched junctions were inserted into the 10 bp chip scaffolds (scaffolds 5, 6, or 7 in [B]; green to blue). WC pairs within the 9 bp chip scaffolds are included for reference (scaffolds 2, 3, or 4 in [B]; pink). Green dashed lines indicate the range observed for the WC pairs in the 10 bp chip scaffolds.

(G) Heatmap depicting the number of 2x2 junctions (with the indicated mismatch in each of the two positions) that fall below lower green dotted line in (F). Mismatch types with black outline and white asterisks have significant enrichment above expected, as in (E) (adjusted p value < 0.05; see STAR Methods).

See also Figure S5.

Internal Loops Can Compensate for Distal Structural Perturbations

Junctions with a broader range of underlying conformational states may be more likely to accommodate formation of tertiary assemblies that would otherwise be too poorly positioned in a rigid context to form the stabilizing tertiary interactions. We observed this type of behavior for more complex internal loops (i.e., 2x2 and 3x3 internal loops). When challenged with the

shorter (9 bp) flow piece, all 10 bp WC chip pieces were significantly destabilized (~2 kcal/mol) compared to the 9 bp WC chip pieces (~0.5 kcal/mol), suggesting these 10 bp pieces cannot easily accommodate the conformational strain of binding the shorter flow piece (Figure 4F, green versus pink violin plots). 1x1 mismatches are similar to WC pairs in this behavior, while some of the 2x2 (30 of 864) and many of the 3x3 junctions (103 of 864) had substantially smaller effects, approaching the minimal effect of the flow piece change for the 9 bp WC chip pieces (Figure 4F). These junctions presumably have readily accessible conformational states that shorten or untwist the helix within the bound tectoRNA, thus compensating for the misalignment of the tertiary contacts imposed by deleting a bp in the binding partner. The 2x2 motifs with this behavior are dominated by a CC mismatch above or below a GG mismatch (Figure 4G), suggesting these types of internal loops allow access to more "compressed" conformational states.

Bulge Conformational Behavior Is Largely Unaffected by Bulge Residue Identity

We systematically explored the attributes of bulged junctions that lead to distinct conformational behavior by comparing the thermodynamic fingerprints of bulged junctions in multiple flanking sequence contexts. Hierarchical clustering of the thermodynamic fingerprints of single bulges (0x1 and 1x0; N = 16 junctions) revealed that the overarching discriminating attribute between junctions was the flanking sequence context, followed by the insertion side, and lastly, the purine-pyrimidine identity of the bulged residue (Figures 5B and 5C). That the identity of the bulged residue mattered very little to differences between bulged motifs was also confirmed by k-means clustering (Figure S6A) and was quantified using PC analysis (Figures 5D and 5E). Unlike single mismatch motifs, single bulges had position-dependent effects along the helix, consistent with bulge-induced helical kinks (Figures 5B, S6B, and S6C).

Analogous to larger internal loops, larger bulges had more complex conformational behavior (see PC analysis in Figures 5F and 5G; Figure S6D), but these profiles were still largely dominated by the contributions of flanking sequence and insertion side, with much smaller contributions of the sequence of the bulged base (Figures 5G and S6D). The absence of strong effects from the identity of the bulged bases suggests that bulged residues are not forming stacking interactions with their adjacent bps, as those interactions would likely impart different conformational behavior dependent on the identity of the bulged base. Thus, our data support the physical model that bulged residues are predominantly extrahelical under the conditions investigated.

Unbiased Classification of Two-Way Junction Conformational Behavior

To generate a comprehensive, unified picture of the energetic behavior of all 1,687 two-way junctions in our library, we carried out unbiased clustering to associate each junction, independent of their sequence or topological class, with other junction sequences that exhibited similar thermodynamic fingerprints (Figure 6A; see STAR Methods). These associations were determined by assigning to each individual junction a set of "neighbors," based on having very similar thermodynamic behavior (see STAR Methods). On average, neighbors were within 0.2 kcal/mol MAD from the target junction thermodynamic fingerprint.

The sets of junction neighbors defined through this analysis were enriched for sequences with the same topology (Figures 6B and 6C), further supporting the idea that the arrangement of unpaired bases is a major factor in determining distinct conformational behavior (Figure 3A above). Certain junction sequences had few neighbors and thus exhibited distinct thermodynamic behavior, implying 3D conformational preferences distinct from other sequences of similar topology (Figures 6D and 6E). Junctions with increasing numbers of non-WC-paired residues in combination with a bulged residue (i.e., 1x2, 1x3, etc.) were enriched for junction sequences that exhibited distinct thermodynamic profiles (Figure 6E). This unbiased clustering also revealed a subset of junctions that clustered with different topological classes than their predicted secondary structure would suggest (e.g., 1x3 motifs that cluster with 2x2 motifs; 2x2 and 3x3 motifs that cluster with bulges; Figure 6B). These junctions presumably have internal interactions that lead to conformational behavior that overlaps with that of other topologies, and these sequences represent targets for future structural studies.

Specific Structural Differences Underlie Thermodynamic Behavior

We next aimed to uncover the relationship between thermodynamic behavior and the underlying structural behavior of junctions. We used our unbiased clustering to relate classes of junctions with common thermodynamic fingerprints to conformations of two-way junctions that have been previously crystallized. These structurally characterized junction sequences (359 of 1,687 total) were extracted from the PDB crystal structure database and ranged from single mismatches and bulges to more complex motifs such as kink turns, right-hand turns, and larger asymmetric bulges (Petrov et al., 2013). Structurally characterized junctions were often among the neighbor sets of other junctions: while only 377 structurally characterized junction sequences were measured, over 1,000 junction sequences had at least one structurally characterized junction in its neighbor set.

We reasoned that the set of structures associated with any neighbor of a junction might provide a reasonable approximation of the distribution of end-to-end distances and orientation changes spanned by that junction, thereby allowing construction of a "stand-in" conformational ensemble. To explore this possibility, we identified a subset of junctions that had among their neighbor sequences at least 10 structurally characterized junctions (148 of the 1,687 junctions), which included 48 junctions with no previous structural characterization. Figure 7A shows the thermodynamic fingerprints of five of these junctions, together with the fingerprints of their neighbors, with the resulting "stand-in" conformational ensemble shown for each junction in Figure 7B. Figure 7C shows the coordinates of each structure projected into the top two "structural PCs" (determined by PC analysis of the six-dimensional structural coordinates of junctions; STAR Methods), illustrating structural features associated with these five thermodynamic fingerprints. We observed variation between the junctions' stand-in ensembles, both in terms of the average structural PCs (i.e., center of contour plots) and their distributions. Certain structures are outliers in this space

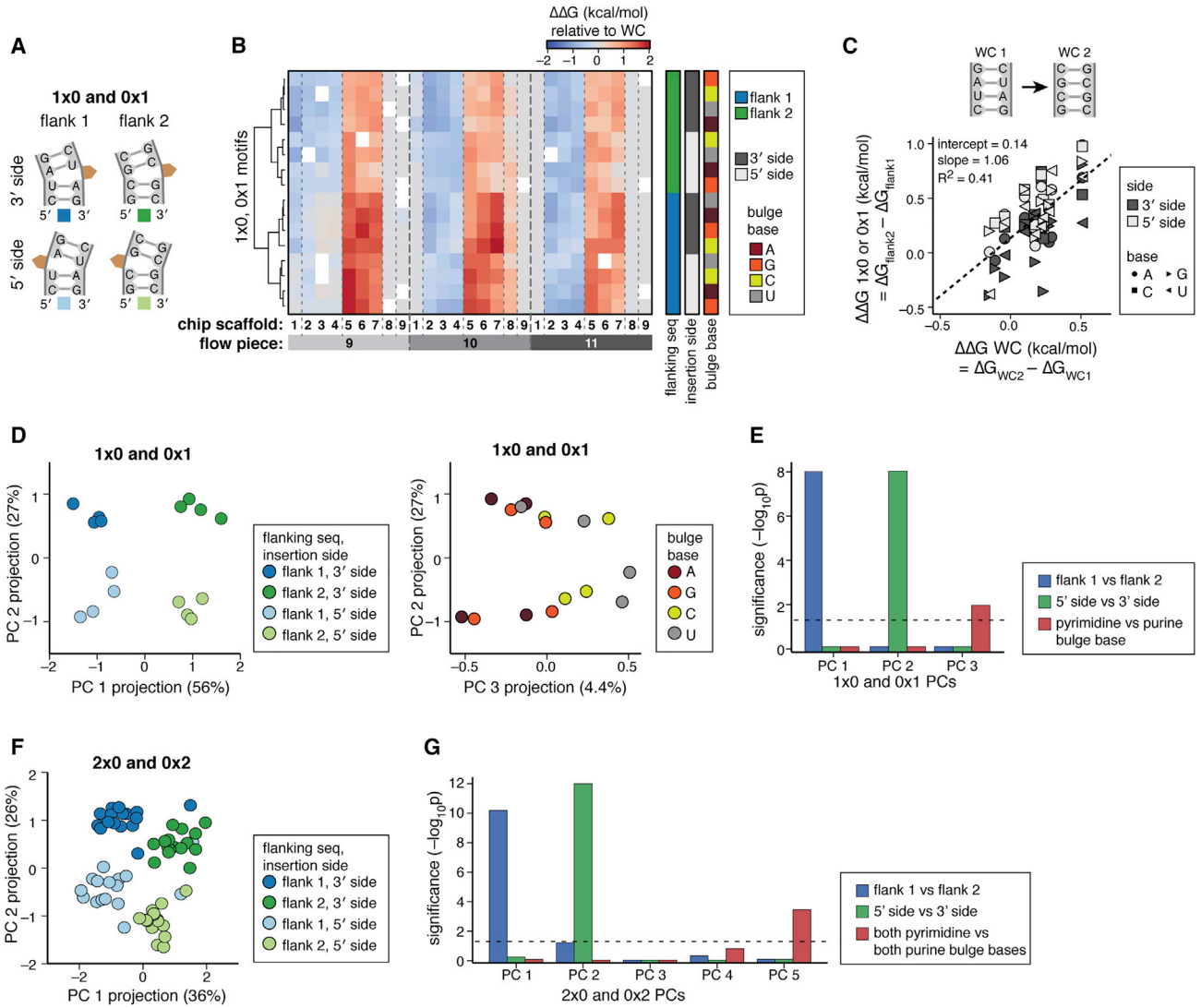


Figure 5. Bulged Junctions Have Independent Contributions of Flanking Sequence, Insertion Side, and Bulge Identity

(A) Schematic of the 16 single-bulge junctions.

(B) Heatmap of the hierarchically clustered thermodynamic fingerprints of the single bulge junctions, relative to the average WC profile. Chip scaffold numbers are defined in Figure 4B. Color bars indicate the attributes of each junction (right).

(C) Scatterplot of the difference in affinity between two WC sequences with versus without an inserted bulge residue, across 11 chip-scaffold/flow-piece contexts with stable binding. Surrounding bps in flank 1 are identical to WC 1, correspondingly for flank 2 and WC 2.

(D) Scatterplots of 0x1 and 1x0 thermodynamic fingerprints projected into the top three PCs. Colors indicate flanking sequence and insertion side (left) or the identity of the bulged residue (right). Percentages indicate the fraction of variance associated with each PC.

(E) Significance (adjusted p value) of the difference between the PC projections of single bulge junctions, divided into two groups based on flanking sequence (blue), insertion side (green), or purine/pyrimidine identity of the bulge (red). Dashed line represents p values of 0.05, values above line are significant.

(F) Scatterplot of 0x2 and 2x0 thermodynamic fingerprints, projected into the top two PCs. PC 2 versus PC 1 is shown. Percentages indicate the amount of variance in each PC; marker colors denote bulged motif attributes as in (A) and (D).

(G) Significance (adjusted p value) of the difference between the PC projections of 2x0 and 0x2 fingerprints, with values divided as in (E) except bulge base identity, which is evaluated between the junctions with both bulged bases being purine or pyrimidine (red).

See also Figure S6.

compared to the structures of thermodynamically related neighbors (e.g., one structure marked with arrow in Figure 7C). These structures may represent rare, higher-energy conformational states of the ensemble that are stabilized in certain structural contexts, underscoring that a single crystallographic conformation may not accurately represent the conformational behavior of a junction (see also below).

To understand how the behaviors of junctions map between these two subspaces—"thermodynamic" and "structural"—we compared the top PCs in each of these spaces. Because each

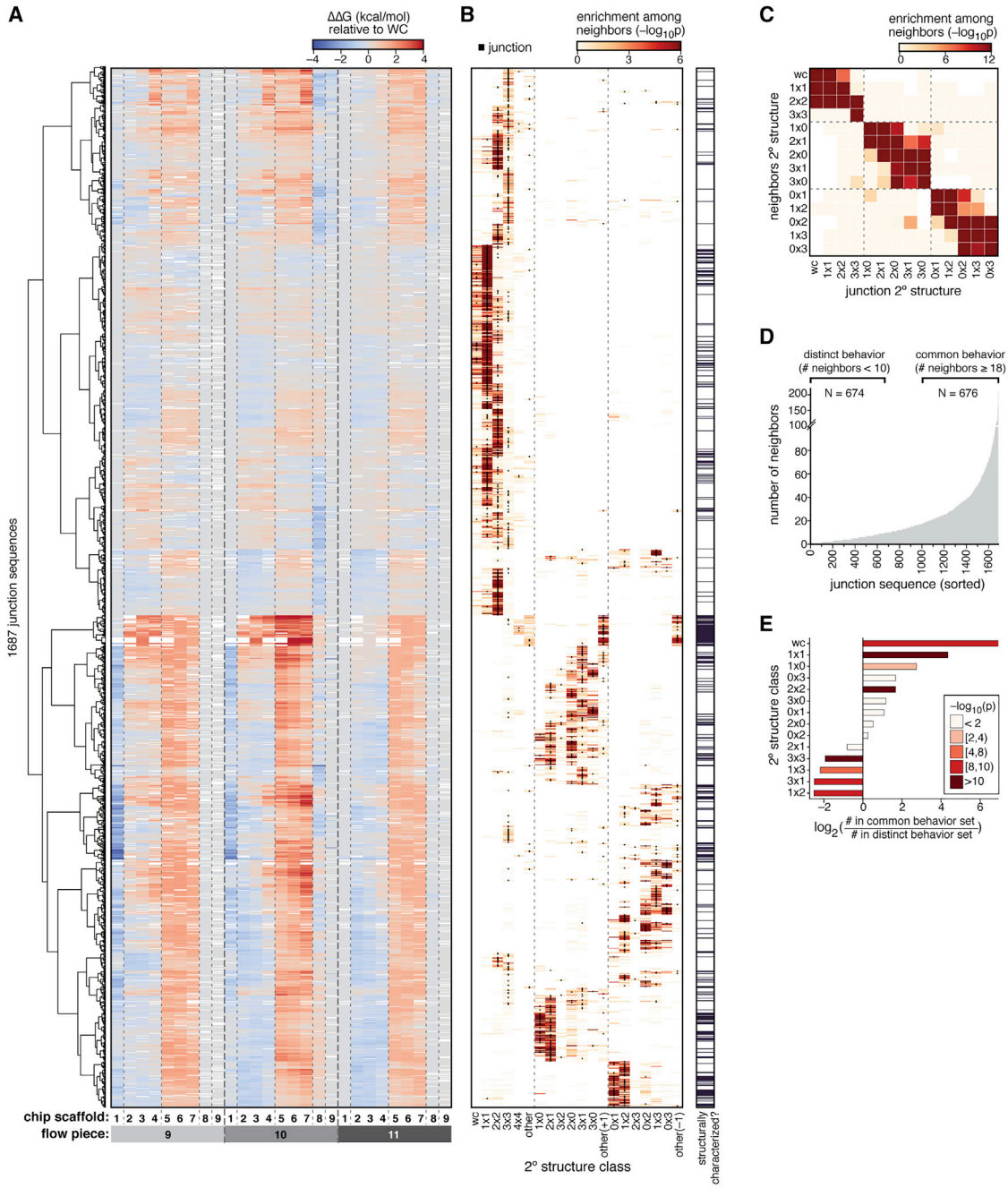


Figure 6. *De Novo* Clustering of Junctions Defines Conformationally Interchangeable Motifs

(A) Heatmap of the clustered thermodynamic fingerprints of 1,687 two-way junctions, relative to average WC fingerprint. Scaffolds are defined in Figure 4B. Projections into the top six PCs were clustered hierarchically to obtain the dendrogram (left) (STAR Methods).

(B) Heatmap indicates the significance of the enrichment for each secondary structure class among its neighbors (STAR Methods); black points overlaid show the secondary structure class of the junction itself. Colorbar indicates whether each junction has been previously structurally characterized (right) (black). Secondary structure classes "other," have more than three non-WC pairs.

(C) Heatmap showing for each secondary structure class (x axis), the enrichment for each other secondary structure class (y axis) among the union of neighbors of members of that class.

(D) The number of neighbor sequences associated with each junction.

(E) For each secondary structure class, bars show the enrichment among the class members for "common" thermodynamic behavior versus "distinct" thermodynamic behavior as defined in (D). Colors of the bars indicate the significance (STAR Methods).

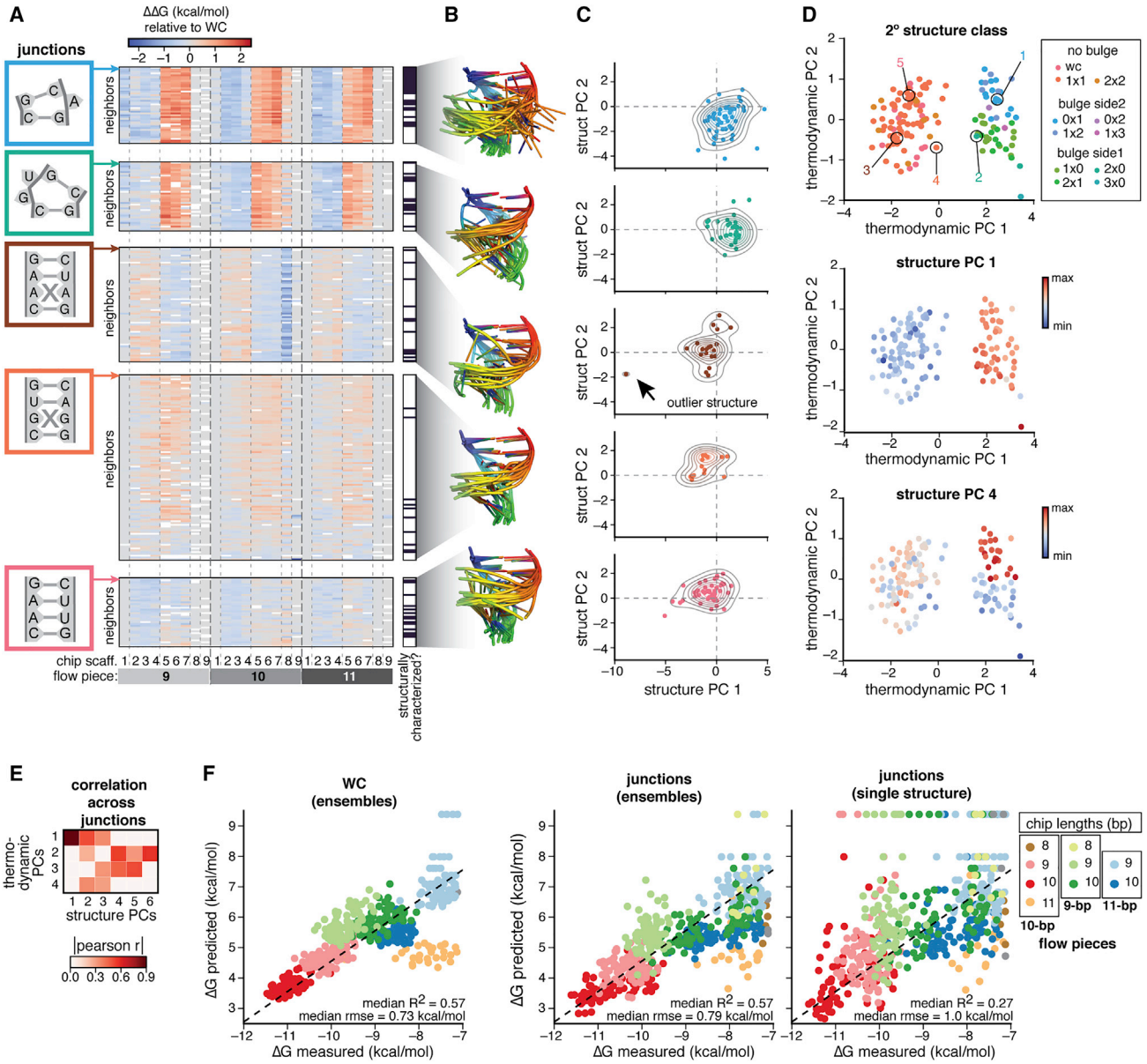


Figure 7. Structurally Characterized Junctions Enable Prediction of Assembly Energetics

(A) Heatmap of thermodynamic fingerprints for five example junction elements (indicated at left), together with the fingerprints of each junction's neighbors. Color bar indicates structurally characterized junctions.

(B) Conformational ensembles generated by aligning the structures of neighbor junction sequences from (A), showing a total of six paired residues.

(C) Projection of end-to-end positions of structural ensembles in (B) into the top two "structural PCs" to reduce the dimensionality from six to two dimensions (STAR Methods and Figure S7A). Contour plots show the kernel-density smoothed distributions.

(D) Scatterplot shows the thermodynamic fingerprints of a subset of junctions projected into two "thermodynamic PCs" (STAR Methods and Figure S7B). Points are colored by secondary structure class (top), or the average value of the structural PC 1 (middle) or PC 4 (bottom) across each junction's structurally characterized neighbors.

(E) Heatmap shows the correlation between values of top four thermodynamic PCs and the top six structural PCs across junction sequences. Any correlations not found to be significant (adjusted $p < 0.05$ after accounting for multiple hypotheses) were set to zero.

(F) Scatterplot of the observed versus predicted affinity of tectoRNA flow/chip contexts containing either WC pairs or two-way junctions. WC bps within the tectoRNA helix are modeled as an ensemble; junctions are modeled using the conformational ensemble derived from grouping the crystallographic structures of the motif and its neighbors ("ensembles") or using simply the crystallographic structure of the motif ("single structure"). Dashed line has a slope equal to 1; intercept is the average difference between observed and predicted across variants.

See also Figure S7.

junction had multiple structures within its ensemble, average PC values across structures were used to characterize "structural" behavior. We found significant correlations between the thermodynamic and structural behavior of junctions (Figures 7D, 7E, and S7C). For example, values in thermodynamic PC 1 are highly correlated to those of structural PC 1 across junctions (Pearson $r = 0.74$; Figure 7D [middle] and Figure 7E); both axes separate bulged from non-bulged junctions. Changes in structural PC 1 correspond to spanning a smaller distance in x- and z-directions and a greater distance in the y-direction, with a corresponding change in rotational angles (Figure S7A); these changes are consistent with "kinking" of the helix. Differences in other structural PCs correspond to other differences in thermodynamic profiles (Figures 7D–7E and S7C); for example, structural PC 4 distinguishes bulged junctions with insertions on one side from the other, corresponding to thermodynamic PC 2 (Figure 7D, top versus bottom panel). Thus, there is a mapping between thermodynamic fingerprints and specific structural behaviors, suggesting that distinct thermodynamic behavior in the tectoRNA system is a readout of transferable structural differences.

Junction Ensembles Predict Thermodynamics of TectoRNA Formation

We next employed a computational model for the tectoRNA binding process to relate the structural descriptions of junctions to their thermodynamic effects. This model was previously used to successfully predict the substantial differences in affinity among tectoRNA variants with different composition of WC bps (Figure 7F, "WC") (Yesselman et al., 2018). Structural ensembles of the constituent elements of the tectoRNA heterodimer were convolved to predict the probability of aligning the second tertiary contact of a tectoRNA once the first contact was formed; this probability provides a measure of the relative affinity for each heteroduplex variant (STAR Methods). We modeled the relative affinity of 30 structurally characterized junctions across each of the flow/chip contexts comprising the thermodynamic fingerprint. Each junction was modeled either by the ensemble of at least 10 structural states associated within its neighbor set or its single crystallographic state (Figure 7F, "ensembles" versus "single structure"). Using the ensemble of structures for each junction, the model predicted observed affinity well, with RMSE values similar to predictions made for variants with only WC bps (median R^2 of 0.57 and RMSE of 0.79 kcal/mol compared to R^2 of 0.57 and RMSE of 0.73 kcal/mol for WC pairs only). In contrast, the single crystallographic conformation of each junction produced worse predictions (median R^2 of 0.27 and RMSE of 1.0 kcal/mol), with many heterodimers observed to bind that were predicted not to bind within our measurable range (Figures 7F and S7D). These results demonstrate that our stand-in ensembles provide a reasonable approximation of the end-to-end distances and orientations exhibited by each junction and thus are expected to enable thermodynamic prediction of tertiary assembly formation.

DISCUSSION

Since the recognition that RNAs form tertiary structures and catalyze reactions (Kruger et al., 1982; Woese et al., 1983), a major goal has been to understand how a primary sequence of RNA encodes the formation of a functional tertiary structure. Here, we took a thermodynamic-centric approach to characterize the sequence-structure relationships of RNA two-way junction elements at high throughput. Our results support the perspective of RNA modularity, albeit in a modified form. In contrast to "structural modularity," in which large structures can be assembled from structures of its individual constituent elements like LEGOs, our results support the concept of "ensemble modularity," in which the likelihood of forming a tertiary structure can be calculated by convolving the structural ensembles of its constituent elements. We found that using the structural ensemble for junction sequences was more predictive than using a single crystallographic conformation (Figures 7F and S7D), supporting the effectiveness of ensemble models and suggesting that RNA elements cannot be treated as static structures when making energetic predictions. This principle of "ensemble modularity" may enable quantitative, energetic models for tertiary structure formation, especially when combined with additional knowledge of tertiary contact thermodynamic stability and potentials for electrostatic repulsion forces (Herschlag et al., 2015).

Generation and refinement of structural ensembles of diverse RNA elements is the next step in using "ensemble modularity" to model RNA tertiary formation, and these data provide a resource to guide future efforts. Most simply, our analysis identified numerous junction sequences that are representative of other junctions, such that high-resolution structural characterization of these junctions will maximize coverage of the conformational landscape of RNA two-way junctions. We determined putative ensembles for several junction elements by grouping crystal structures of similarly behaving junctions. Encouragingly, these are ensembles predictive of tectoRNA binding thermodynamics, though it is possible that these ensembles may be limited in their ability to predict junctions in the context of tightly packed RNA structures or in complex with proteins. We envision that these ensembles will be improved over time, perhaps by using structure prediction algorithms in combination with the high-dimensional constraints provided by our data, e.g., Frank et al. (2009). Finally, using cryoelectron microscopy (cryo-EM) to obtain high-throughput, direct ensemble visualization of a variety of RNA elements may be possible in the future; our technology and the conceptual framework presented here will allow incisive tests of such future ensembles (Zhang et al., 2018).

Full accounting of the behavior of complex RNAs will require energetic descriptions for the entire complement of RNA elements. Thermodynamic fingerprints can be readily generated for other RNA elements that do not substantially change the geometry of forming both tertiary contacts in the tectoRNA system, including three- and four-way junctions and variants of the TL/TLR tertiary contacts used here. Further, the effects on the conformational preferences of any of these elements due to ligand binding or ionic conditions can be studied with thermodynamic fingerprint analysis, with potential impact on aptamer design and small-molecule binding to biological and therapeutically relevant RNA tertiary structures. However, examination of tertiary elements such as kissing loops that impose conformational preferences very different than the TL/TLRs will require the development of different

host systems than the tectoRNA to probe their conformational behavior within a stable tertiary assembly. Development of different host scaffolds would also enable probing different regions of conformational space, possibly allowing broader characterization of conformations within junction conformational ensembles.

Our data illustrate that many motifs can compensate "at a distance" for structural perturbations; for example, destabilization from a decrease in the length of the chip-piece helix by a single bp can be compensated by introduction of a bulged residue or by shortening the flow piece (Figure 2B). These compensating variants represent multiple solutions to forming a stable tertiary assembly without affecting the tertiary contact itself. This observation has implications for our understanding of functional RNA structures and their evolution. Previously, evolutionary covariation of residues in RNA secondary and tertiary structures has been used to identify residues that directly interact—like those involved in WC bps and in tertiary contact interactions—as these residues have mutually evolved to conserve the stable secondary or tertiary assembly (Weinreb et al., 2016). Compensation at a distance may also be common in the evolution of tertiary structure but diffusely distributed throughout structures and thus difficult to discern through covariation analyses.

Consideration of the ability of a junction to compensate at a distance for a structural perturbation may prove useful for the engineering and design of structured RNAs. Junctions that are ineffective at compensating for other perturbations likely have more homogeneous ensembles and thus are more likely to be structurally modular across contexts. This class of junctions may be the most useful for building RNAs with specified conformations. In contrast, conformational heterogeneity is important for engineering dynamic behaviors (Bailor et al., 2010), and our data suggest that more complex junctions often exhibit a greater breadth of conformational flexibility (i.e., 3x3 junctions in Figure 4F) that helps tertiary contact formation in a broader range of structural contexts. These larger internal loops may blunt the perturbative effects of other mutations, insertions, or deletions within an RNA stem, thereby enhancing the evolvability of RNA tertiary structure (Wagner and Altenberg, 1996). This flexibility also suggests a design principle for the selection of functional RNA aptamers in which the starting pool of variants contains larger internal loops to more readily attain a desired structure.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Library design
- METHOD DETAILS
  - Assembly and sequencing of library
  - Fluorescent labeling of RNA flow pieces
  - Imaging station setup
  - Generation of RNA on the sequencing flow cell
  - Measuring dissociation constants on chip
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Processing sequencing data
  - Data processing and image fitting
  - Fluorescence normalization
  - Binding curve fitting
  - Off-rate fits and photobleaching correction
  - Evaluating significant effects
  - Accounting for inter-experimental error
  - Combining experimental replicates
  - Data filtering
  - Handling of missing data
  - Calculation of mean absolute deviation
  - K-means clustering of single mismatch motifs
  - Significance of motif attribute enrichment
  - Unbiased clustering of all two-way junctions
  - Extracting structural coordinate of junctions
  - Prediction of thermodynamics with RNAMake-ΔG
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information includes seven figures, four tables, and one data file and can be found with this article online at https://doi.org/10.1016/j.cell.2018.05.038.

ACKNOWLEDGMENTS

We thank Curtis Layton and Johan Andreasson for developing, building, and maintaining the imaging station; Greenleaf and Herschlag lab members for reagents and critical feedback; and Olke Uhlenbeck and Luc Jaeger for helpful feedback and discussions. This work was supported by the National Institute of Health (grant number P01GM066275 to D.H., GM111990 and P50HG007735 to W.J.G., GM100953 and GM122579 to R.D.), and the Beckman Foundation. W.J.G. acknowledges support as a Chan-Zuckerberg Investigator. S.K.D. was supported in part by the Stanford Biophysics training grant (T32 GM008294) and by the NSF Graduate Research Fellowship. N.B. was supported in part by the NSF Graduate Research Fellowship. J.D.Y. was supported by the Ruth L. Kirschstein National Research Service Award Postdoctoral Fellowships GM112294.

AUTHOR CONTRIBUTIONS

Conceptualization, S.K.D., N.B., J.D.Y., R.D., D.H., W.J.G.; Formal Analysis, S.K.D., N.B., J.D.Y.; Software—Thermodynamic prediction simulation model, J.D.Y.; Writing—Original draft, S.K.D., N.B., D.H., W.J.G.; Writing—Final draft, S.K.D., N.B., J.D.Y., R.D., D.H., WJ.G.; Project Administration, D.H., W.J.G.; Funding Acquisition, R.D., D.H., W.J.G.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: November 30, 2017
Revised: March 7, 2018
Accepted: May 15, 2018
Published: June 28, 2018

REFERENCES

Altschul, S.F., and Gish, W. (1996). Local alignment statistics. In Methods Enzymol. (Elsevier), pp. 460–480.

Bailor, M.H., Sun, X., and Al-Hashimi, H.M. (2010). Topology links RNA secondary structure with global conformation, dynamics, and adaptation. Science 327, 202–206.

Bisaria, N., Greenfeld, M., Limouse, C., Mabuchi, H., and Herschlag, D. (2017). Quantitative tests of a reconstitution model for RNA folding thermodynamics and kinetics. Proc. Natl. Acad. Sci. USA *114*, E7688–E7696.

Brion, P., and Westhof, E. (1997). Hierarchy and dynamics of RNA folding. Annu. Rev. Biophys. Biomol. Struct. *26*, 113–137.

Buenrostro, J.D., Araya, C.L., Chircus, L.M., Layton, C.J., Chang, H.Y., Snyder, M.P., and Greenleaf, W.J. (2014). Quantitative analysis of RNA-protein interactions on a massively parallel array reveals biophysical and evolutionary landscapes. Nat. Biotechnol. *32*, 562–568.

Chauhan, S., and Woodson, S.A. (2008). Tertiary interactions determine the accuracy of RNA folding. J. Am. Chem. Soc. *130*, 1296–1303.

Chu, V.B., Lipfert, J., Bai, Y., Pande, V.S., Doniach, S., and Herschlag, D. (2009). Do conformational biases of simple helical junctions influence RNA folding stability and specificity? RNA *15*, 2195–2205.

Costa, M., and Michel, F. (1995). Frequent use of the same tertiary motif by self-folding RNAs. EMBO J. *14*, 1276–1285.

Costa, M., and Michel, F. (1997). Rules for RNA recognition of GNRA tetraloops deduced by in vitro selection: comparison with in vivo evolution. EMBO J. *16*, 3289–3302.

Cruz, J.A., and Westhof, E. (2009). The dynamic landscapes of RNA architecture. Cell *136*, 604–609.

Frank, A.T., Stelzer, A.C., Al-Hashimi, H.M., and Andricioaei, I. (2009). Constructing RNA dynamical ensembles by combining MD and motionally decoupled NMR RDCs: new insights into RNA dynamics and adaptive ligand recognition. Nucleic Acids Res. *37*, 3670–3679.

Geary, C., Baudrey, S., and Jaeger, L. (2008). Comprehensive features of natural and in vitro selected GNRA tetraloop-binding receptors. Nucleic Acids Res. *36*, 1138–1152.

Herschlag, D., Allred, B.E., and Gowrishankar, S. (2015). From static to dynamic: the need for structural ensembles and a predictive model of RNA folding and function. Curr. Opin. Struct. Biol. *30*, 125–133.

Jaeger, L., and Leontis, N.B. (2000). Tecto-RNA: One-Dimensional Self-Assembly through Tertiary Interactions. Angew. Chem. Int. Ed. Engl. *39*, 2521–2524.

Klein, D.J., Schmeing, T.M., Moore, P.B., and Steitz, T.A. (2001). The kink-turn: a new RNA secondary structure motif. EMBO J. *20*, 4214–4221.

Kruger, K., Grabowski, P.J., Zaug, A.J., Sands, J., Gottschling, D.E., and Cech, T.R. (1982). Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena. Cell *31*, 147–157.

Lorenz, R., Bernhart, S.H., Höner Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F., and Hofacker, I.L. (2011). ViennaRNA Package 2.0. Algorithms Mol. Biol. *6*, 26.

Moore, M.J. (2005). From birth to death: the complex lives of eukaryotic mRNAs. Science *309*, 1514–1518.

Murchie, A.I., Thomson, J.B., Walter, F., and Lilley, D.M. (1998). Folding of the hairpin ribozyme in its natural conformation achieves close physical proximity of the loops. Mol. Cell *1*, 873–881.

Mustoe, A.M., Brooks, C.L., 3rd, and Al-Hashimi, H.M. (2014). Topological constraints are major determinants of tRNA tertiary structure and dynamics and provide basis for tertiary folding cooperativity. Nucleic Acids Res. *42*, 11792–11804.

Needleman, S.B., and Wunsch, C.D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. *48*, 443–453.

Noller, H.F. (2005). RNA structure: reading the ribosome. Science *309*, 1508–1514.

Petrov, A.I., Zirbel, C.L., and Leontis, N.B. (2013). Automated classification of RNA 3D motifs and the RNA 3D Motif Atlas. RNA *19*, 1327–1340.

Qin, P.Z., Butcher, S.E., Feigon, J., and Hubbell, W.L. (2001). Quantitative analysis of the isolated GAAA tetraloop/receptor interaction in solution: a site-directed spin labeling study. Biochemistry *40*, 6929–6936.

Salmon, L., Yang, S., and Al-Hashimi, H.M. (2014). Advances in the determination of nucleic acid conformational ensembles. Annu. Rev. Phys. Chem. *65*, 293–316.

She, R., Chakravarty, A.K., Layton, C.J., Chircus, L.M., Andreasson, J.O.L., Damaraju, N., McMahon, P.L., Buenrostro, J.D., Jarosz, D.F., and Greenleaf, W.J. (2017). Comprehensive and quantitative mapping of RNA-protein interactions across a transcribed eukaryotic genome. Proc. Natl. Acad. Sci. USA *114*, 3619–3624.

Shi, X., Beauchamp, K.A., Harbury, P.B., and Herschlag, D. (2014). From a structural average to the conformational ensemble of a DNA bulge. Proc. Natl. Acad. Sci. USA *111*, E1473–E1480.

Šidák, Z. (1967). Rectangular Confidence Regions for the Means of Multivariate Normal Distributions. J. Am. Stat. Assoc. *62*, 626–633.

Strulson, C.A., Boyer, J.A., Whitman, E.E., and Bevilacqua, P.C. (2014). Molecular crowders and cosolutes promote folding cooperativity of RNA under physiological ionic conditions. RNA *20*, 331–347.

Thomson, J.B., and Lilley, D.M. (1999). The influence of junction conformation on RNA cleavage by the hairpin ribozyme in its natural junction form. RNA *5*, 180–187.

Tinoco, I., Jr., and Bustamante, C. (1999). How RNA folds. J. Mol. Biol. *293*, 271–281.

Turner, D.H., Sugimoto, N., and Freier, S.M. (1988). RNA structure prediction. Annu. Rev. Biophys. Biophys. Chem. *17*, 167–192.

Wagner, G.P., and Altenberg, L. (1996). Perspective: complex adaptations and the evolution of evolvability. Evolution *50*, 967–976.

Watkins, A.M., Geniesse, C., Kladwang, W., Zakrevsky, P., Jaeger, L., and Das, R. (2018). Blind prediction of noncanonical RNA structure at atomic accuracy. Sci. Adv. *4*. Published online May 25, 2018. https://doi.org/10.1126/sciadv.aar5316.

Weinreb, C., Riesselman, A.J., Ingraham, J.B., Gross, T., Sander, C., and Marks, D.S. (2016). 3D RNA and Functional Interactions from Evolutionary Couplings. Cell *165*, 963–975.

Woese, C.R., Gutell, R., Gupta, R., and Noller, H.F. (1983). Detailed analysis of the higher-order structure of 16S-like ribosomal ribonucleic acids. Microbiol. Rev. *47*, 621–669.

Wu, L., Chai, D., Fraser, M.E., and Zimmerly, S. (2012). Structural variation and uniformity among tetraloop-receptor interactions and other loop-helix interactions in RNA crystal structures. PLoS ONE *7*, e49225.

Yesselman, J.D., Denny, S.K., Bisaria, N., Herschlag, D., Greenleaf, W.J., and Das, R. (2018). RNA tertiary structure energetics predicted by an ensemble model of the RNA double helix. bioRxiv. https://doi.org/10.1101/341107.

Zhang, K., Keane, S.C., Su, Z., Irobalieva, R.N., Chen, M., Van, V., Sciandra, C.A., Marchant, J., Heng, X., Schmid, M.F., et al. (2018). Structure of the 30 kDa HIV-1 RNA Dimerization Signal by a Hybrid Cryo-EM, NMR, and Molecular Dynamics Approach. Structure *26*, 490–498.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Chemicals, Peptides, and Recombinant Proteins		
Phusion Hot Start Flex DNA Polymerase	NEB	M0535
Phusion HF buffer	NEB	B0518
SYBR green	ThermoFisher	S-11494
dNTPs	NEB	N0447
NTP set	NEB	N0450
5x SSC buffer	ThermoFisher	15557036
Klenow 3′-5′ exo(–)	NEB	M0212
NEB Buffer 2	NEB	B7002
E. coli RNA Polymerase, Holoenzyme	NEB	M0551
Streptavidin	PROzyme	SA10
D-Biotin	ThermoFisher	B20656
Yeast tRNAs	ThermoFisher	AM7119
Critical Commercial Assays		
QIAquick PCR Purification Kit	QIAGEN	28106
NEBNext High-Fidelity 2X PCR Master Mix	NEB	M0541
PhiX Control V3	Illumina	FC-110-3001
Qubit RNA HS Assay Kit	ThermoFisher	Q32852
Deposited Data		
Example sequencing library variant	This paper	https://benchling.com/s/QhoiCB/edit
Nucleotide substitution scoring matrix	BLAST	ftp://ftp.ncbi.nih.gov/blast/matrices/NUC.4.4
RNA Structure Atlas v 1.45	(Petrov et al., 2013)	http://rna.bgsu.edu/rna3dhub/nrlist/release/1.45
Oligonucleotides		
DNA sequences of chip piece tectoRNA variants, see Table S1	CustomArray, this paper	N/A
Oligos used for construction of library, see Table S2	IDT DNA, this paper	N/A
Oligos used for *in situ* transcription, see Table S3	IDT DNA, this paper	N/A
RNA sequences of flow piece tectoRNA binding partner, see Table S4	IDT DNA, this paper	N/A
Software and Algorithms		
RNAFold (v 2.1.8)	(Lorenz et al., 2011)	https://www.tbi.univie.ac.at/RNA/RNAfold.1.html
Sequencing data processing	This paper	N/A
Image data quantification	This paper	N/A
Dissociation constant fitting	This paper	N/A
Thermodynamic fingerprint neighbor determination	This paper	N/A
Prediction of tectoRNA assembly formation (RNAMake-ΔG)	This paper, (Yesselman et al., 2018)	N/A

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, William Greenleaf (wjg@stanford.edu).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Library design

TectoRNAs are composed of a helical segment flanked by a hairpin loop and by a tertiary contact receptor (Jaeger and Leontis, 2000; Geary et al., 2008). Sequence variants of each of these elements were designed to derive the sequences of each tectoRNA variant (sequences can be found in Table S1, an overview of designs is found in Figures 1D and S1B). Each tectoRNA variant was assembled by defining its receptor type, loop sequence, helix sequence, helix length, incorporated junction sequence, and the position of the incorporated junction.

METHOD DETAILS

Assembly and sequencing of library

Designed library variants together with common priming regions on the 5′ and 3′ sides were synthesized into DNA by Custom Array (Bothell, WA; Figure S1D). Each sequence was ordered in duplicate using 92,000 oligo-scale synthesis. The longest length of any library member was 113 bp, and sequences shorter than this were appended with polyA at the 3′ end to give a uniform length. The synthesized oligo pool was amplified using internal primers to enrich for full-length library variants (oligopool_left and oligopool_right; Table S2). The PCR reaction consisted of: 1:400 dilution of the synthesized oligo pool, 200 nM of each primer, 200 μM dNTPs, 3% DMSO, 1x Phusion HF buffer, 0.01U/μl of HS Phusion (NEB M0535). The reaction proceeded for 9 cycles of 98°C for 10 s, 62°C for 30 s, and 72°C for 30 s. Reaction mixtures were purified using QIAquick PCR Purification Kit (QIAGEN 28106) to remove primers and proteins and eluted into 20 μL of elution buffer.

After this initial amplification, the library was amplified with distinct primers to bring in sequences compatible with Illumina sequencing as well as a unique molecular identifier (UMI) barcode consisting of a 16 nt randomer. This five-piece assembly PCR included two outside primers and two adaptor sequences. The PCR reaction consisted of 1 μL of the previous reaction, 137 nM of outside primers (short_C and short_D; Table S2), 3.84 nM of the adaptor sequences (C1_R1_BC_RNAP and D_Read2; Table S2), 200 uM dNTPs, 3% DMSO, 1x Phusion HF buffer, and 0.02U/μl of Phusion Hot Start Flex enzyme (NEB). The reaction proceeded for 14 cycles of 98°C for 10 s, 63°C for 30 s, and 72°C for 30 s. Reactions were purified using with QIAquick PCR Purification Kit.

Finally, the library was bottlenecked to reduce the representation of unique molecular identifiers to ~700,000 molecules as follows. To obtain accurate estimate of the concentration, the library was diluted 1:5000 (in 0.1% Tween20). This dilution was quantified against a standard library of PhiX (Illumina, Hayward, CA). PhiX was first diluted to 25 pM in 0.1% Tween20, and subsequently diluted serially 2-fold in 0.1% Tween20, to form a standard curve from 25 pM to 0.2 pM. The diluted library and the PhiX standard were each amplified in a qPCR assay to determine their relative cycle thresholds (CT). The qPCR reaction consisted of 1.25 μM primers (short_C, short_D; Table S2), 0.6x SYBR green (ThermoFisher S-11494), and 1x NEBNext Master Mix (NEB, M0541), and the cycling conditions were 98°C for 10 s, 63°C for 30 s, and 72°C for 60 s.

The concentration of the diluted library was quantified using the CT values of the PhiX standard (mean of triplicate CT values), and the volume associated with 700,000 molecules was added to a new PCR reaction (1.25 μM short_C and short_D, 1x NEBNext PCR Master Mix), and amplified for 21 cycles, using the same cycling conditions. The bottlenecked, amplified library was purified with QIAquick PCR Purification Kit. Finally, the library was sequenced on an Illumina Miseq instrument. The bottlenecked library represented between 10%–30% of the total sequencing chip, and the remaining sequences consisted of high-complexity genomic libraries. The Miseq was sequenced in three cycles: 75 bases in read 1, 75 bases in read 2, and an 8 bp i7 index read, to obtain demultiplexed, paired-end sequencing information.

See https://benchling.com/s/QhoiCB/edit for annotated sequence and primers of a representative member of the library.

Fluorescent labeling of RNA flow pieces

Sequences of flow pieces are given in Table S4. RNA oligos modified with 5′-Amino Modifier C6 were ordered from Integrated DNA Technologies (Coralville, IA), HPLC-purified, and then labeled with an NHS-ester conjugated Cy3b dye. Reactions were ethanol-precipitated overnight at –20°C, gel purified using a denaturing polyacrylamide gel (8% PAGE, 8 M urea, 1x TBE: 89 mM Tris-HCl, 89 mM Boric Acid, pH 7.4, 2 mM sodium EDTA), then eluted in water after three freeze-thaw cycles. To reduce aggregation on the chip surface, stock flow piece solutions were spun through a 50K Amicon filter (Amicon UFC505008) two times and collected on a 3K Amicon filter (Amicon UFC500308). Flow pieces were quantified after purification using Qubit RNA high sensitivity kit (Thermofisher).

Imaging station setup

An imaging station was built from a combination of custom-designed parts and pieces from disassembled Illumina genome analyzer IIx, as based on (Buenrostro et al., 2014) and modified as in (She et al., 2017). Briefly, the custom parts included a fluidics adaptor designed to interface between the Illumina Miseq chip and the fluidics pump, a temperature control system that maintained the temperature of the flow cell, a laser control circuit, and filters for the different imaging channels. Two channels were employed: the "red" channel used the 660 nm laser and 664 nm long pass filter (Semrock) and the "green" channel used the 530 nm laser

and a 590 (104) nm band pass filter (Semrock). All images were taken with 400 ms exposure times at 200 mW fiber input laser power. Focusing was achieved by sequential adjustment of the z position and re-imaging of the four corners of the flow cell; the adjusted z-positions for each of the four corners was then fit to a plane. This plane then gave the z-position for each of the 16 tiles.

Generation of RNA on the sequencing flow cell

RNA was generated by *in situ* transcription of a DNA library arrayed on an Illumina Miseq sequencing flow cell (Buenrostro et al., 2014; She et al., 2017). All steps were run using an in-house imaging station. Unless otherwise stated, washing or buffer exchange was done by flowing 250 μL volume through the chip at 100 μl/min.

Regeneration of double-stranded DNA

Post-sequencing, the chip was washed with Cleavage buffer (100 mM Tris-HCl, 125 mM NaCl, 0.05% Tween20, 100 mM TCEP, pH 7.4) to remove residual fluorescence from the reversible terminators used in the sequencing reaction at 60°C for 5 min. The strand of DNA not covalently attached to the surface of the chip was removed by washing in 100% formamide at 55°C.

The resulting single-stranded DNA fragments were converted to double-stranded DNA to provide a substrate for RNA generation (Figure S1D(i)). Double-stranded DNA generation was achieved by first hybridizing a DNA oligo with a 5′ biotin to the chip: the chip was incubated with 500 nM of the oligo Biotin_D_Read2 (Table S3) in Hybridization buffer (5x SSC buffer (ThermoFisher 15557036), 5 mM EDTA, 0.05% Tween20) for 15 min at 60°C, subsequently the temperature was lowered to 40°C for another 10 min. The chip was washed with Annealing buffer (1x SSC buffer, 7 mM MgCl2, 0.01% Tween20), and then in Klenow buffer (1x NEB buffer 2 (NEB B7002S), 250 M each dNTP, 0.01% Tween20). Extension of the primer was achieved by applying one line volume (150 μl) of Klenow buffer with 0.1 U/μl Klenow fragment (3′-5′ exo(–) (NEB M0212)), followed by half of a line volume of Klenow buffer. This process minimized the amount of enzyme within the fluidics line during the incubation period. The Klenow enzyme was incubated in the flow cell at 37°C for 30 min, then was washed with Hybridization buffer.

In the event that double-stranded DNA generation was not 100% efficient, we annealed an excess of complementary oligos to the stall sequence lacking a fluorophore (Dark_stall; Table S3) in order to completely block any single-stranded segments of DNA within this region. The chip was incubated with 500 nM of Dark_stall in Hybridization buffer at 37°C for 10 min. The chip was then washed in Annealing buffer, and a second hybridization was done with 500 nM of Dark_stall in Annealing buffer at 37°C for 10 min. Subsequently the chip was again washed with Annealing buffer. To ensure that none of the stall sequence remained single stranded, 500 nM of the Fluorescent_stall (Table S2) in Annealing buffer was incubated in the chip at 37°C for 10 min, and the chip was washed with Annealing buffer. The chip was then imaged to ensure minimal fluorescence on the DNA clusters.

Generation of RNA

Before beginning transcription, the flow cell was incubated with 1 μM streptavidin (PROzyme SA10) in Annealing buffer, which bound the biotinylated primer from which the second strand of DNA was extended (Figure S1D(i)). After washing in Annealing buffer, the flow cell was incubated with 5 μM biotin (Thermofisher B20656) in Annealing buffer to saturate remaining biotin binding sites within each streptavidin tetramer (Figure S1D(ii)). To transcribe RNA from the dsDNA clusters, *E. coli* RNA polymerase holoenzyme (RNAP; NEB M0551) was allowed to initiate, but not significantly extend, the RNA transcript, thereby limiting the number of RNAP molecules that could initiate on each DNA molecule, as in (Buenrostro et al., 2014; She et al., 2017). RNAP extension was limited by the absence of CTP in the initiation condition. The flow cell was washed with Initiation buffer (2.5 μM each of ATP, GTP, and UTP in R-reaction buffer, which consists of 20 mM Tris-HCl pH 8.0, 7 mM $MgCl_2$, 20 mM NaCl, 0.1% BME, 0.1 mM EDTA, 1.5% Glycerol, 0.02 mg/ml BSA, and 0.01% Tween20). Transcription was initiated by applying one line volume (150 μl) of 0.06 U/μl RNAP in Initiation buffer, followed by half of a line volume of Initiation buffer. The RNAP was allowed to bind and initiate transcription for 20 min at 37°C, then the chip was washed with Initiation buffer to remove excess unbound RNAP in solution. The chip was then washed with Extension buffer (R-reaction buffer, with 1 mM of each NTP). Transcription was allowed to extend for 10 min at 37°C. Ultimately, the stalled polymerase displays the nascent transcript on the surface of the sequencing chip (Figure S1D(iii)).

DNA oligos were annealed to the nascent transcript to both assess the efficiency of transcription and to block ssRNA common to all transcribed molecules and not part of the specified variable region (Figures 1C and S1D). To perform this annealing, the chip was incubated with the Fluorescent_stall which hybridizes to the transcribed RNA and Dark_read2, each at 500 nM in Annealing buffer, for 10 min at 37°C. Finally, the flow cell was washed with Binding buffer (89 mM Tris-Borate, pH 8.0, 30 mM $MgCl_2$, 0.01 mg/ml yeast tRNAs (ThermoFisher Scientific AM7119), 0.01% Tween20). The temperature of the chip was lowered to 20°C, then imaged. This set of images served as the quantification for transcription efficiency (red channel), as well as the baseline fluorescence with no labeled flow piece in solution (green channel). However, since this image showed so little fluorescence, registering this image to our data was not possible in any experiment, so this second image was not quantified. Instead, we relied on fitting to obtain the baseline fluorescence per cluster in the absence of ligand (see Data processing and image fitting for registration and quantification description, and Binding curve fitting and estimation of fmax for low-affinity variants for description of defining baseline and maximal fluorescence values).

Measuring dissociation constants on chip

Affinity was determined for each tectoRNA cluster on the surface of the flow cell by applying increasing concentrations of labeled tectoRNA "flow piece" in solution (Figures 1C and 1E) and measuring the amount of fluorescence at each cluster at each concentration. The flow piece was diluted to 10 μM in water, denatured at 95°C for 1 minute, then refolded on ice for 2 min. 5x Binding buffer

and water was then added to bring the final concentration of the flow piece to 2 μM in 1x Binding buffer. A total of eight dilutions were made by 3-fold serial dilutions in Binding buffer, for a lowest concentration of 0.91 nM flow piece. Experiments were carried out at 22°C.

Per experiment, each dilution of the flow piece was applied to the flow cell in increasing concentration, and incubated for 3 hr, 2 hr, 1 hr, 45 min, 30 min, 20 min, 20 min, and 20 min, for the 8 concentration points, respectively. These variable equilibration times allowed measurements at low concentrations of the time to equilibrate. After equilibration, the fluorescence in the red and green channels was imaged across all the tiles.

Immediately following the equilibrium experiment, off-rates were measured by applying an unlabeled flow piece at high concentration (2 μM) to the chip at a fast flow rate of 150 μl/min. All tiles were imaged sequentially for a total of 20 images per tile (~40 s elapsed between each image), then imaging rate was decreased to 5 min between each imaging round, for a total of 40 images per tile.

QUANTIFICATION AND STATISTICAL ANALYSIS

Processing sequencing data

Processing sequencing data

Sequencing data from the Illumina Miseq was processed to extract the tile and coordinates of each sequenced cluster. Because our sequencing library of tectoRNA chip piece variants was only 10%–30% of the total sequencing library, each sequenced cluster was evaluated for whether it had an RNAP initiation site. This was achieved by doing a Needleman-Wunsch global alignment of the RNAP initiation sequence (TTTATGCTATAATTATTTCATGTAGTAAGGAGGTTGTATGGAAGACGTTCCTGGATCC) to the read1 sequence of each cluster, using the using the Nuc.4.4 scoring matrix (ftp://ftp.ncbi.nih.gov/blast/matrices/NUC.4.4) (Needleman and Wunsch, 1970). The score obtained from Needleman-Wunsch was converted to a p value as described in (Altschul and Gish, 1996), and sequences with p value < 10^{-4} were deemed to have the RNAP initiation sequence. While all clusters are fit during the fitting of the images (described below), this subset of sequences were used to determine the cross-correlation between the images and the sequencing data. In addition, only these clusters were fit in the downstream data analysis (i.e., for determination of K_d and k_{off}).

Assignment of library variants to UMIs

As described in (Buenrostro et al., 2014), incorporation of a 16-nt UMI into the library can ultimately minimize the effect of sequencing error on our interpretations. During assembly of the sequencing library, this 16-nt UMI was incorporated by PCR into each DNA fragment, and the library was bottlenecked to increase the representation of each individual UMI within the overall library. Each UMI was sequenced deeply to form a consensus sequence of the associated tectoRNA sequence that was then compared to our designed library (~20-fold coverage of each UMI; per-base consensus). This consensus sequence identifies the molecular identity of the library variant with much lower frequency of errors than would occur across the entire tectoRNA sequencing read.

UMIs were evaluated based on the fraction of representative sequences that matched the consensus sequence. If the number of sequences matching the consensus could be explained by chance, or if the UMI was too short (< 12 nt), the UMI was not included in subsequent analysis (and any clusters associated with this UMI were similarly not analyzed). The chances of matching the consensus sequence were assumed to be binomial, where a success is a match and a failure is a discrepancy, and the rate of success under the null model is 25%. This filter removed UMIs associated with diverse unrelated sequences, or with relatively few reads per UMI. The consensus sequence of each filtered UMI was then evaluated for matching each designed library variants to attribute clusters (i.e., with sequenced UMI's) to particular library variant designs.

Data processing and image fitting

Images taken during RNA array experiments were mapped to sequencing data from the Illumina Miseq. First, sequencing data were processed to extract the tile and coordinates of each sequenced cluster. To match each sequence to its location on our imaging station, the sequencing data were cross-correlated to the images in an iterative fashion to eventually map coordinates to the images at sub-pixel resolution as in (She et al., 2017). Once these locations were determined, every cluster was fit to a 2D Gaussian to quantify its fluorescence (Buenrostro et al., 2014).

Fluorescence normalization

To reduce inter-cluster variation in maximum fluorescence, we normalized the amount of tectoRNA flow piece bound at a given cluster (measured in the green channel) by the total amount of transcribed RNA in that cluster (measured in the red channel). To prevent artifacts caused by dividing by small numbers, the fluorescence of the transcribed RNA was required to be greater than the first percentile of the set of fluorescence values in the red channel, and any fluorescence measurement below this threshold was set to the threshold value.

Binding curve fitting

Equilibrium measurements at varying concentration of tectoRNA flow piece were used to determine free energy of binding (ΔG) to each molecular variant in the library. This fitting procedure was carried out stepwise to allow robust fitting across a range of affinities: 1) Binding isotherms were first fit to the fluorescence values of each single cluster. 2) Distributions of fit parameters f_{min} and

f_{max} were determined from these initial fits. 3) Fit refinement of binding isotherms was carried out using these estimates of f_{min} and f_{max} distributions. The last step minimized the attribution of changes in ΔG to spurious changes in f_{min} or f_{max}, especially for variants that didn't fully saturate (i.e., that had unknown f_{max}) or that had evident binding at the lowest concentration of flow piece (i.e., had unknown f_{min}). These three steps are described in detail below.

1) Single-cluster fitting

Initially, fluorescence quantified from each single cluster was fit to a binding isotherm, according to the equation:

$$f(x) = f_{\min} + f_{\max}\frac{x}{x + \exp\left(\frac{\Delta G}{RT}\right)},$$

where f is the normalized fluorescence, f_{min}, f_{max}, and ΔG are free parameters, x is the concentration, R is the gas constant, and T is the temperature in Kelvin. Least-squares fitting was carried out using the python package lmfit (v 0.8.3).

2) Estimating distributions of fmax and fmin

After completing the initial single cluster fits, we determined initial values for f_{min} and f_{max} for each molecular variant by taking the median across single cluster fits associated with each variant. We found systematic biases in the fit parameters, especially for molecular variants that had high K_d values relative to our highest concentration of tectoRNA flow piece (Figure S2C). For these variants, the flow piece did not saturate the chip-piece binding sites during the course of the binding experiment, resulting in poor estimates of f_{max} and ΔG. We hypothesized that we could use the f_{max} values of variants that did saturate to obtain accurate ΔG for those variants that did not saturate. Therefore, we implemented a method (described below) to generate an estimate of the distribution of f_{max} values (based on f_{max} values of variants that fit well) that allowed us to estimate ΔG for the lower affinity variants.

The f_{max} distribution was assumed to follow a gamma distribution with a fixed mean for the entire experiment with a standard deviation dependent on the number of clusters per molecular variant. As the number of clusters per variant increases, the standard error on the mean of our estimate for f_{max} decreases, and thus as the distribution of f_{max} values narrows, we can obtain more precise estimates of ΔG.

To find the mean and standard deviation of the distribution of f_{max}, the fit parameters of individual sequence variants were filtered to obtain variants that both (1) came close enough to saturated binding to allow robust estimation of f_{max} (i.e., was a "tight binder"), and (2) fit the binding equation well (i.e., was a "good fitter"). These characteristics were evaluated as follows: (1) If the median fit value for ΔG for a variant implied that its clusters were at least 95% bound at the highest concentration of flow piece, i.e., if $\Delta G \leq RT\log(0.05x_{final})$, this variant satisfied the "tight binder" criterion. (2) Using the output from the least-squares fitting step, each single-cluster fit was considered a "good" fit if it satisfied the following criteria: a) The coefficient of determination (R^2) was greater than 0.5, b) the standard error on ΔG was less than 1 kcal/mol, and c) the standard error on f_{max} was less than f_{max}. For each molecular variant, we determined whether the number of "good" fits associated with its single clusters could have happened by chance, assuming a rate of success under a null model of 25% (empirically determined). If the null hypothesis could be rejected ($p < 0.01$), this molecular variant was assumed to satisfy the "good fit" criterion. In addition to selecting variants that fit well, this process selects for variants that were measured many times.

The set of molecular variants that were both good fitters and tight binders were used to find the f_{max} distribution. The global mean f_{max} was obtained by fitting f_{max} values to a gamma distribution and obtaining the mean of the distribution, μ_{global} (Figure S2D). To determine the relationship between standard deviation of f_{max} and the number of clusters per variant n, the distribution of initial f_{max} with n clusters was fit to a gamma distribution for each n with at least 10 values for initial f_{max}. Both the mean and standard deviation were allowed to float during this fit process. The relationship between the standard deviation and the number of measurements, $\sigma(n)$, was fit to an analytical function: $\sigma(n) = (a/\sqrt{\sigma}) + b$, where a and b are free parameters (Figure S2E).

The distribution of f_{max} for each molecular variant with n clusters per variant is then the gamma distribution G:

$$G(f_{\max}, a_n, \theta_n) = \left(\frac{f_{\max}}{\theta_n}\right)^{a_n - 1}\frac{\exp(-f_{\max}/\theta_n)}{\Gamma(a_n)}$$

where $a_n = (\mu_{global}/\sigma(n))^2$ and $\theta_n = (\sigma(n)^2/\mu_{global})$. This distribution depends only on the number of clusters per variant, n.

For experiments without a zero-concentration image, some of the variants showed systematic relationships between f_{min} and ΔG, especially very stable variants with significant binding at the lowest concentration point (Figure S2C). To adjust for this, a global value for f_{min} was applied to all variants. This value was obtained by selecting variants that were poor binders (less than 50% bound at the highest protein concentration), and taking the median fluorescence per variant across single clusters at the lowest concentration point, and then finding the median of this set of values.

3) Fit refinement by fitting binding isotherms with fmax distribution

To enforce the value of f_{min} and the distribution of f_{max}, the binding isotherm of each molecular variant was refined as follows. Each molecular variant was assessed based on whether it likely achieved saturation or not. If the median fluorescence across the single clusters of that variant did not exceed the lower bound of the 95% confidence interval on f_{max} (given the distribution of f_{max} and the number of clusters per variant n), this variant was assumed to not have achieved saturation.

For all variants, fitting was carried out iteratively by resampling the clusters associated with each variant (with replacement). Resampling and fitting multiple times allowed (1) estimation of the error on the fit values based on the resulting distribution and (2) definition of f_{max} values for those variants that did not achieve saturation. For each resampling, the median fluorescence for each concentration point was obtained, and that set of values was fit to a binding isotherm, where f_{min} was fixed at the global value described above. This resampling was repeated 100 times to form a 95% confidence interval on each of the fit parameters (f_{max} and ΔG). However, for variants that did not achieve saturation by the above definition, f_{max} was not allowed to float: instead, 100 values were sampled from the f_{max} distribution (using n as the number of clusters associated with that variant). For each iteration of the resampling, f_{max} was set to one of these 100 values, thereby enforcing the estimated distribution of f_{max}. The median fit ΔG obtained from the initial single cluster fits was used as the initial value in the least-squares fitting of each fitting iteration. For variants where f_{max} was allowed to float, the median fit f_{max} was used as the initial value.

Testing of binding curve fitting method

We applied our fit method that used a constrained estimated distribution of f_{max} to variants that did not achieve saturation in the first five concentrations of the binding isotherm series, but did achieve saturation using all eight of the concentrations in the series (Figure S2F). The ΔG values obtained using this subset of concentrations and enforcing the estimated distribution of f_{max} recapitulated with high accuracy the ΔG values obtained using all eight concentrations of the binding series (Figure S2F), suggesting this method provides an accurate estimates of ΔG values for variants that did not achieve saturation.

This fit refinement procedure affected the affinity especially for clusters that had outlying f_{max} or f_{min} values in the initial fit (Figure S2G). In general, variants with low f_{max} were affected the most, and most often the re-fit K_d was greater than the highest measured concentration, as expected.

To further test this method, a set of ''background variants'' was obtained that were sampled from clusters on the chip that do not have RNAP initiation sites and therefore should not exhibit any binding to the tectoRNA flow piece. Each background variant was assigned to a set of background clusters such that the final distribution of clusters/variant recapitulated that of our library. The binding series fluorescence of these variants was then fit exactly as described above, except the f_{max} distribution was estimated using only library members (i.e., those with RNAP initiation sites). The fit refinement procedure greatly increased the fit ΔG values relative to the initial ΔG values (obtained using the median of the single cluster fits of each variant, which allowed f_{max} and f_{min} to float; Figure S2H). Without the fit refinement procedure, many of these background variants would falsely have been attributed ΔG values similar to those of tectoRNA variants with intermediate affinity, but with aberrantly low values for f_{max}. The values of ΔG for this set of non-binders were used to determine the upper bound of the measurable range of affinity (set to Kd of 5000 nM, or ΔG of –7.1). This upper bound was smaller than >99% of the affinity values determined for this set of background variants.

Off-rate fits and photobleaching correction

Per cluster off-rates were derived by fitting single clusters to an exponential decay following dilution, assuming a constant fractional photobleaching in every image taken.

$$f(t,m) = f_{\min} + (f_{\max} - f_{\min})\exp(-k_{off}t)\alpha^m$$

In this case, f_{min}, f_{max}, and k_{off} are free parameters, t is the time (in seconds) at which each image was taken, α is the photobleaching rate, and m is the number of times each image was taken, starting with 1. The photobleaching rate was determined by sequential imaging of a stable interaction between a subset of chip pieces that formed kissing loop structure with the flow piece, and was found to be 0.9924 per image (i.e., 0.76% of the total fluorescence in each cluster was lost every time an image was taken).

The fit parameters of molecular variants were taken as the median of the fit parameters across the single clusters of each variant. The 95% confidence intervals on each variant were found by bootstrapping the fit parameters of each single cluster.

Off-rate measurements per variant were assumed to be reliable if the f_{max} was greater than 0.6 (60% of expected normalized fluorescence) and the error on $RT\ \log(k_{off})$ was less than 0.5 kcal/mol (95% confidence interval). The threshold on f_{max} ensured significant binding was observed at the beginning of the time series.

Evaluating significant effects

To evaluate significance of deviations compared to measurement error, a false discovery approach was used. Each deviation was converted into a z-score, by subtracting the average deviation, and dividing by the error on the deviation. If there was no significant deviation from the overall effect, the distribution of z-scores should be normally distributed with zero mean and standard deviation of one. Given this null distribution, a two-tailed false discovery rate was calculated for each z-score threshold. First, an estimate of the number of false discoveries was made by finding the fraction of measurements more extreme than that z-score threshold for a standard normal distribution ($2\mathrm{CDF}(-|z|)$; CDF = cumulative distribution function of Gaussian distribution), and multiplying by the number of tests, N. The total number of discoveries was the number of false discoveries (above) plus the number of z-scores whose absolute values were greater than or equal to that threshold. False discovery rate (FDR) was then the number of false discoveries over the total number of discoveries. The number of measurements identified as different at a false discovery rate of 0.1 were reported as the number ''significantly'' different from the average effect.

Accounting for inter-experimental error

Comparing replicate experiments, we noticed that our error estimation internal to each experiment (generated from bootstrapping the fit parameters across single-cluster measurements, "intra-experiment error") slightly underestimated the statistical variability observed between the two experiments, even given a small overall offset in the ΔG estimates between the experiments (i.e., from slight differences in estimates of ligand concentrations between experiments). About 2% of the variants could not be explained by this overall offset, and this proportion increases when looking at variants with low "intra-experiment error" (e.g., ~10% for variants with "intra-experiment error" less than 0.025 kcal/mol (standard error)). These data suggested that our "intra-experiment error" estimates were systematically slightly deflated compared to actual (i.e., inter-experimental) error, especially when the error estimate is extremely small.

To quantify this systematic deflation in error estimates across experiments, we first calculated the z-scores of the deviation between the two replicate experiments, where the z-score for each variant is the difference in measured ΔG between the two replicate experiments (i.e., ΔΔG), less the overall offset between the two replicate experiments ($\Delta\Delta G_{avg}$), and divided by the standard error on ΔΔG (i.e., $\sigma_{\Delta\Delta G}$): $z = (\Delta\Delta G - \Delta\Delta G_{avg}/\sigma_{\Delta\Delta G})$, where the standard error on ΔΔG is the combined quadrature error on each of the intra-experimental error estimates: $\sigma_{\Delta\Delta G} = \sqrt{\sigma^2_{\Delta G_1} + \sigma^2_{\Delta G_1}}$. If all of the deviations between the replicate experiments could be accounted for by the overall offset and the intra-experiment error, these scores should be Gaussian distributed with zero mean and unit standard deviation. In contrast, in the case where the replicate measurements differed more than could be explained by the intra-experiment error, the z-score distribution should have standard deviation > 1. To understand the relationship between the estimated intra-experiment error and the inter-experiment deviations, z-scores were divided into 100 equally populated bins based on the combined intra-experiment error, i.e., $\sigma_{\Delta\Delta G}$. For all variants within each bin, the z-scores were calculated and fit to a Gaussian distribution. The relationship between the standard deviation of the z-scores (Σ_z) and the combined standard error of that bin, σ, was empirically determined to follow a power law: $\sum_z(\sigma) = A\sigma^k$. Free parameters *A* and *k* were fit using least-squares regression (A = 0.744, k = –0.25).

Error estimates in these and subsequent experiments were scaled by the output of this function, i.e., the standard error on the fit ΔG values (σ) was multiplied by $\sum_z(\sigma)$ to obtain the scaled standard error for each molecular variant. This error scaling reduced the number of variants that were significantly different from the overall offset between replicate measurements from 4% to less than 0.02% (Figure S2A). Comparison to a different replicate experiment similarly reduced the number of variants that were significantly different from the overall offset from 4% to 0.4%, confirming that this error scaling is generally applicable. For the vast majority of measurements (> 90%), this process slightly increased our uncertainty in our estimation of ΔG, on average by 0.06 kcal/mol. Error estimates are shown in Figures 1G and S2B, separated by ΔG and the number of clusters per variant. This slight inflation of error allowed confident comparison of measured deviations that exceed the inter-experimental error.

Combining experimental replicates

The reported values for the 10-bp flow piece are combined across two independent replicate experiments (Figure 1G). The overall offset between the experiments was subtracted off the second replicate experiment. For each variant, the two measurements (ΔG_1 and ΔG_2), each with error (σ_1 and σ_2), were combined by taking the weighted average of the two values, where the weights are the inverse squared error on each measurement, i.e.: $\Delta G_{comb} = ((\Delta G_1/\sigma_1^2) + (\Delta G_2/\sigma_2^2))((1/\sigma_1^2) + (1/\sigma_2^2))^{-1}$. The combined error is then: $\sigma_{comb} = ((1/\sigma_1^2) + (1/\sigma_2^2))^{-1}$.

Data filtering

Affinity measurements were not included in the final dataset if they had low representation, if the error on the affinity measurement was large, or if they were predicted to have misfolded secondary structure. Specifically, low representation was defined as fewer than 5 measurements made for that molecular variant per chip. Measurements with errors estimated to be greater than 0.5 kcal/mol (95% confidence interval) were also removed from analyses.

Secondary structure misfolding was assessed for tectoRNA chip pieces using RNAfold (2.1.8) (Lorenz et al., 2011). Specifically, the ensemble free energy of folding was assessed at 20°C of the chip piece sequence (command: "RNAfold -p0 -T20"). The ensemble free energy of folding into secondary structures with the loop and receptor formed was also assessed using a constraint ("RNAfold -p0 -C -T20"). The constraint required the first 6 bps below and the two bps above the receptor to be formed as well as for the loop to be unpaired. Measurements were filtered if the difference in the ensemble free energy of the unconstrained sequence and the constrained sequence was greater than 0.5 kcal/mol.

For the global analysis of two-way junctions (Figure 6), 5% of the measurements were removed by these filters, with the majority of these missing measurements due to low representation (85%).

Handling of missing data

Analyses could be missing affinity measurements in several cases: for example, if a certain motif embedded in a particular tectoRNA scaffold had low representation among the library members or if embedding a motif within that scaffold destabilized the secondary structure of the tectoRNA (see above). For clustering or PC analysis, missing data had to be interpolated in order to make comprehensive comparisons (i.e., Figures 4, 5, and 6). When comparing relatively similar motifs (such as mismatch motifs or similar bulges in

Figures 4 and 5), missing data were interpolated by the median affinity measured for other motifs in the same chip-scaffold/flow-piece context. When comparing more disparately behaving motifs, we employed a method to find, for any junction sequences with missing measurements, the 20 most similar thermodynamic fingerprints that were within 0.2 kcal/mol MAD from that profile, based only on measured contexts. Missing data were then interpolated, as above, by the median affinity measured of these related motifs in the missing chip-scaffold/flow-piece contexts. While clustering was carried out on datasets with interpolated data, all plots of individual data points show only uninterpolated data.

Calculation of mean absolute deviation

The deviation of individual junction sequences from the average profile of their secondary structure class (as in Figure 3C) was calculated by finding the mean absolute deviation (MAD) between each sequence and its class average profile using only chip-scaffold/flow-piece contexts in which detectable binding was observed.

K-means clustering of single mismatch motifs

Thermodynamic fingerprints of the 112 individual 1x1 junctions and WC elements were clustered with k-means clustering (see Figure 4C). Any missing values were assigned to the median value across junctions in that context. PC analysis was performed, and the projections into the top 6 PCs were subsequently clustered with k-means clustering (k = 7). Different numbers of clusters were also evaluated (i.e., see Figure S5A).

Significance of motif attribute enrichment

Determining overrepresentation of a motif attribute among a subset of motifs

In multiple cases, motifs were divided into classes, and the significance of enrichment for a given motif attribute (e.g., mismatch type or secondary structure class) was evaluated for each class (i.e., see Figures 4E, 4G, 6B, 6C, and 6E). To perform this analysis, the null model was that a given motif attribute was distributed equally across all classes. Fisher's exact test was then used to determine if the number of motifs with that motif attribute in a given cluster was enriched relative to the null expectation (one-sided p value). This test was repeated for each class and for each of the motif attributes. All p values were then corrected for multiple hypotheses using the Holm-Sidak method (Šidák, 1967).

Determining significance of quantitative differences between two sets of motifs

In other cases, we asked whether there was a significant different between the values characterizing two sets of motifs that were divided based on their motif attribute (e.g., Figures 5E and 5G). In this case, p values are calculated using a t test between the values in the two sets, and corrected for multiple hypothesis testing (Šidák, 1967).

Unbiased clustering of all two-way junctions

To generate the hierarchical clustering shown in Figure 6A, thermodynamic fingerprints were decomposed into the top six principal components, each of which corresponded to a distinct mode of behavior. The remaining PCs were associated with < 10% of the variance and could not be easily attributed to a physical perturbation. Projections into each of these six PCs were standardized by dividing by the standard deviations of the projections in each PC to equally weight each of these modes of behavior. Motifs were clustered by the Euclidean distance between each of these vectors (hierarchical ward clustering; Figure 6A).

To determine the neighbors of a junction sequence, we initially found all other junction sequences within Euclidean distance of 2 (based on the top six standardized PC projections of their thermodynamic fingerprints as above). For each of these candidate neighbors, a threshold was imposed on the deviation in affinity in each flow piece/chip scaffold context, which was three-fold the error on the free energy of binding of the reference sequence in that structural context. If the candidate junction profile did not deviate by more than the threshold in any structural context, it was included in the neighbor set.

Extracting structural coordinate of junctions

Structures associated with two-way junction sequences within the non-redundant set of structurally characterized RNAs (Petrov et al., 2013) were aligned by their top bp (i.e., closest to the hairpin loop). Flanking bp steps were modeled into the structure such that each assembled structure (i.e., junction and flanking bps) spanned a total of six bps, to facilitate comparison of junctions of different sizes. Flanking bp structures were modeled by the most common configuration of base-pair steps extracted from the same set of structurally characterized RNAs (Yesselman et al., 2018). The end-to-end distance and orientation between this top bp and the structure's bottom bp were extracted in terms of the translational coordinates (x, y, and z) and the rotational coordinates (α, β, γ), relative to that of a WC element spanning the same number of bps. Figure 7B shows the aligned structures of many junction elements.

We initially focused on a subset of junctions that had at least 10 structurally characterized neighbors (178 junctions). Several of these junctions were large (i.e., contained many unpaired residues) and often highly destabilized in any tectoRNA context, making analysis challenging; for this reason junctions with more than three non-WC paired residues or more than three unpaired, bulged residues were not included in subsequent analysis (148 junctions after applying this filter). Most of these sequences had structures associated with them, but one-third (48 junctions) did not. We performed structural PC analysis on the set of structures associated with any neighbor of these 148 junctions (378 structures associated with 194 unique junction sequences; Figure S7A). Values of each

coordinate were divided by their standard deviation across all structures, in order to be able to compare translational and rotational differences, before PC analysis. The PC loadings are shown in Figure S7A, after multiplication by the standard deviation of each coordinate to give each value in terms of physical quantities (i.e., distances in Angstrom or rotations in radians).

Prediction of thermodynamics with RNAMake-ΔG

To make a stand-in structural ensemble for a given junction sequence, the set of neighbors associated with that junction were obtained. Neighboring junction sequences were filtered to ensure that the neighbor and the reference junction replaced the same number of bps in the tectoRNA helix. This step is necessary to predict tectoRNA affinity, which requires each element to be modular from the others (i.e., the set of structures corresponding to the junction element are independent of the set of structures associated with the surrounding helical elements). Any structures associated with these filtered neighbor sequences were extracted and aligned as above to form an ensemble of structural states (Petrov et al., 2013). Each structure within the ensemble was weighted equally.

Simulation of the relative affinity of tectoRNA variants was performed as described in (Yesselman et al., 2018). In brief, structural ensembles were obtained for each of the constituent elements of the tectoRNA. These ensembles could consist of one or multiple structural states, and structural states themselves were weighted according to their Boltzmann probability. Ensembles of WC base-pair steps were derived from crystal structures of RNA helices (Petrov et al., 2013). Ensembles of the two bound tertiary contacts consisted only of a single conformation each. In the case of the GAAA-11ntR, the structure was isolated from the crystal structure of the P4-P6 domain of the *Tetrahymena* ribozyme (PDB: 1GID). The other tertiary contact (GGAA-R1) has not been characterized crystallographically and was modeled using Rosetta stepwise Monte Carlo modeling (Watkins et al., 2018).

The tectoRNA flow and chip pieces were assembled into a partially bound state with one tertiary contact pre-formed. The un-formed tertiary contact was not explicitly modeled, and thus structures contain a break between the helix emanating from the tetraloop of this contact and the target bp this tetraloop interacts with in the bound conformation. During the simulation, elements were randomly chosen to be replaced by a different member of this element's structural ensemble. If the new ensemble member had lower energy (based on Boltzmann weights of the ensemble members), then it was accepted. If not, it was accepted with some probability given by the metropolis criteria. Whenever the conformation of a single element was modified, this change was propagated to the entire rest of the modeled tectoRNA structure.

For each step of the simulation, the tectoRNA was determined to be "bound" or "unbound" based on the distance between the emanating bp of the un-formed tertiary contact and the position it would exist in in the bound structure. This distance was defined as the translational distance (in Angstroms) plus the rotational difference between the two bps (in radians). Structures with a distance score of < 5 were called as "bound." The simulation was run for >1 million iterations. Affinity was assumed to be proportional to $-k_B T \log(N_{bound}/N_{unbound})$.

DATA AND SOFTWARE AVAILABILITY

The thermodynamic measurements generated in this paper are available for download (Table S1).

Custom software for processing sequence data, for determination of the dissociation constants, and for obtaining thermodynamically related neighbors are also available for download (Data S1).

Cell

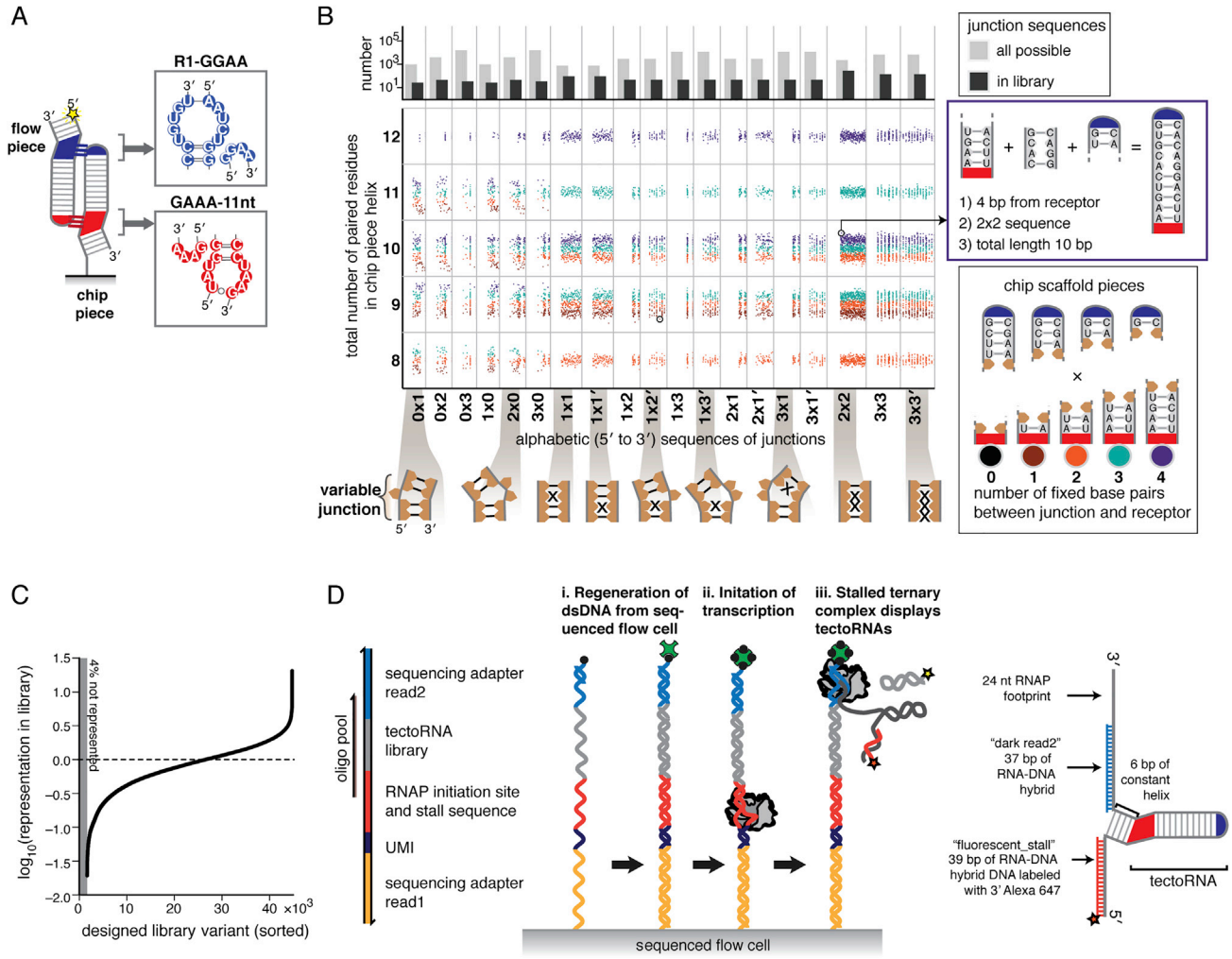


Figure S1. Design and Construction of TectoRNA Library for *In Situ* Transcription on the Surface of an Illumina Sequencer, Related to Figure 1

(A) Sequence and secondary structure of the two tetraloop-tetraloop receptors in the tectoRNA complex, GGAA-R1 and GAAA-11nt.

(B) The junction library was designed by inserting junctions into tectoRNA chip scaffolds. Plot shows inserted junction elements, separated by secondary structure class (e.g., 0x1, 0x2, etc.) along the x axis. Junction elements are inserted into chip scaffolds of varying lengths, represented by the location along the y axis. Each class of junction has many possible primary sequence variants, with location along the x axis indicating the alphabetic position of that sequence (from 5′ to 3′) among all possible sequences with that secondary structure, with the colors representing the location of the inserted junction element, defined on bottom right of figure. Positional jitter was added along the y axis to enable visualization. Subsets of junctions within each class were incorporated in the designed library (quantified by bar graph, top). The width of each secondary structure class in the plot is proportional to the log of the number of motifs incorporated within that category. This sublibrary comprised ~16,000 distinct molecules. Not shown are junction sequences from previously structurally characterized library or control loop/receptor mutants.

(C) Representation of individual chip piece variants in the assembled sequencing library used for measurements.

(D) On-chip workflow for generation of displayed tectoRNAs on the surface of the sequencing flow cell. (Far right) Configuration of on-chip RNA during the assay. Two DNA oligonucleotides are annealed to single-stranded regions of the transcribed RNA that are common to all library variants (dark_read2 and fluorescent_stall; see STAR Methods); one of these oligos is labeled at its 3′ end for quantification of transcribed RNA.

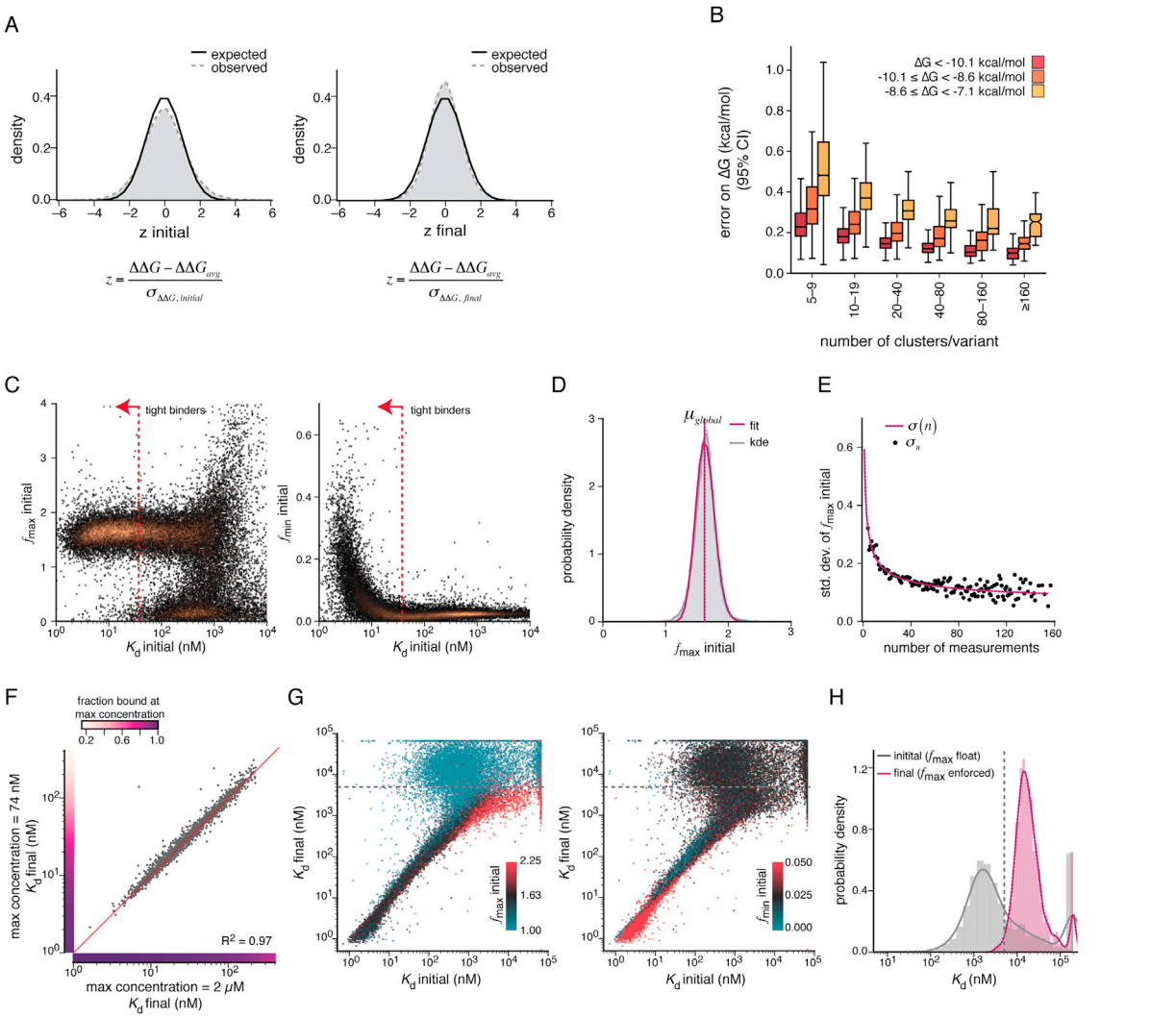


Figure S2. Binding Isotherms Were Fit Constraining Distributions of f_{min} and f_{max}, Related to STAR Methods

(A) Deviation between replicate experiments accounted for by intra-experimental error estimates. Distribution of z-scores giving the difference between replicate experiments, using either the initial (left) or final (right) intra-experimental error estimates, where initial estimates come from bootstrapping across clusters associated with each variant, and final estimates come from scaling these initial values to account for additional inter-experimental error (see STAR Methods). For each variant, z-scores were obtained by finding the difference in measured ΔG between experiments (ΔΔG), relative to the average difference in ΔG between experiments ($\Delta\Delta G_{avg}$), and dividing by the standard error on ΔΔG ($\sigma_{\Delta\Delta G}$), which is the two intra-experimental error estimates combined in quadrature. Black line indicates the expected distribution of z-scores if the difference between the two replicate experiments could be entirely accounted for by the average offset and the intra-experimental error estimates (see STAR Methods).

(B) Distribution of error estimates on ΔG for variants with different number of clusters and with different ΔG values. The estimate of ΔG increases in precision with increasing numbers of measurements and with higher affinity.

(C) Initial fit values for K_d, f_{min}, and f_{max} show interdependent relationships that are likely artifactual, where f_{max} is the binding saturation level (plotted versus K_d; left), and f_{min} is the fluorescence in the absence of any flow piece in solution (plotted versus K_d; right). Initial values are the median across the unconstrained single cluster fits associated with each library variant. These initial values for K_d, f_{min}, and f_{max} were subsequently refined to remove these relationships, as described in the STAR Methods. This refinement especially targeted variants that did not have saturated binding at the highest concentration of the flow piece. Dotted lines indicate the maximum K_d associated with 95% saturation of binding at the highest concentration of the flow piece, and thus demarcate "tight binders."

(D and E) A global distribution for f_{max} enabled fit refinement. The mean (μ_{global}) (D) and standard deviation as a function of number of clusters $\sigma(n)$ (E) of f_{max} values are shown. These values were subsequently used to refine binding curve fitting for variants that did not achieve saturation (see STAR Methods). Initial f_{max} values shown here were obtained from tight binders (i.e., see (C)) and good fitters (see STAR Methods). (D) Histogram of initial f_{max}'s and kernel density estimate ("kde") shown in gray; best-fit gamma distribution ("fit") shown in purple. (E) Scatterplot relating the standard deviation (σ) of initial f_{max} values for variants with n number of observed clusters. The fit relationship between σ and n (purple line) is indicated.

(F) Scatterplot compares the final fit K_d for a set of variants that when using all eight concentration points (max concentration = 2 μM), the f_{max} is well-defined, but when only using the first five concentration points (max concentration = 74 nM), saturation is not achieved and thus f_{max} is not well-defined. Enforcing the

(legend continued on next page)

f_{max} distribution during fit refinement (see STAR Methods) allows accurate estimation of K_d even when saturation is not achieved. Colorbars along the axes indicate the relationship between K_d and the expected fraction bound at the maximum concentration, when using five versus eight concentrations points.
(G) Scatterplots comparing the initial and final K_d values per variant. The final fits are obtained after the fit refinement procedure that enforces the distribution of f_{max} for variants that do not achieve saturation. Dotted lines indicate the cutoff for measurable affinity. Colors indicate the initial f_{max} (left) or f_{min} (right).
(H) Histogram of initial and final K_d for a set of ''background'' variants, in gray and purple, respectively. Fits of these variants reflect non-specific binding or accumulation of background fluorescence, as the clusters associated with these background variants did not have an RNAP initiation site, thus are not expected to generate RNA. The vertical dashed line indicates the measurable range of affinity. Prior to fit refinement, (i.e., without enforcing f_{max}), many of these background variants fall within the measurable range of affinity because they are fit to unrealistic f_{max} values.

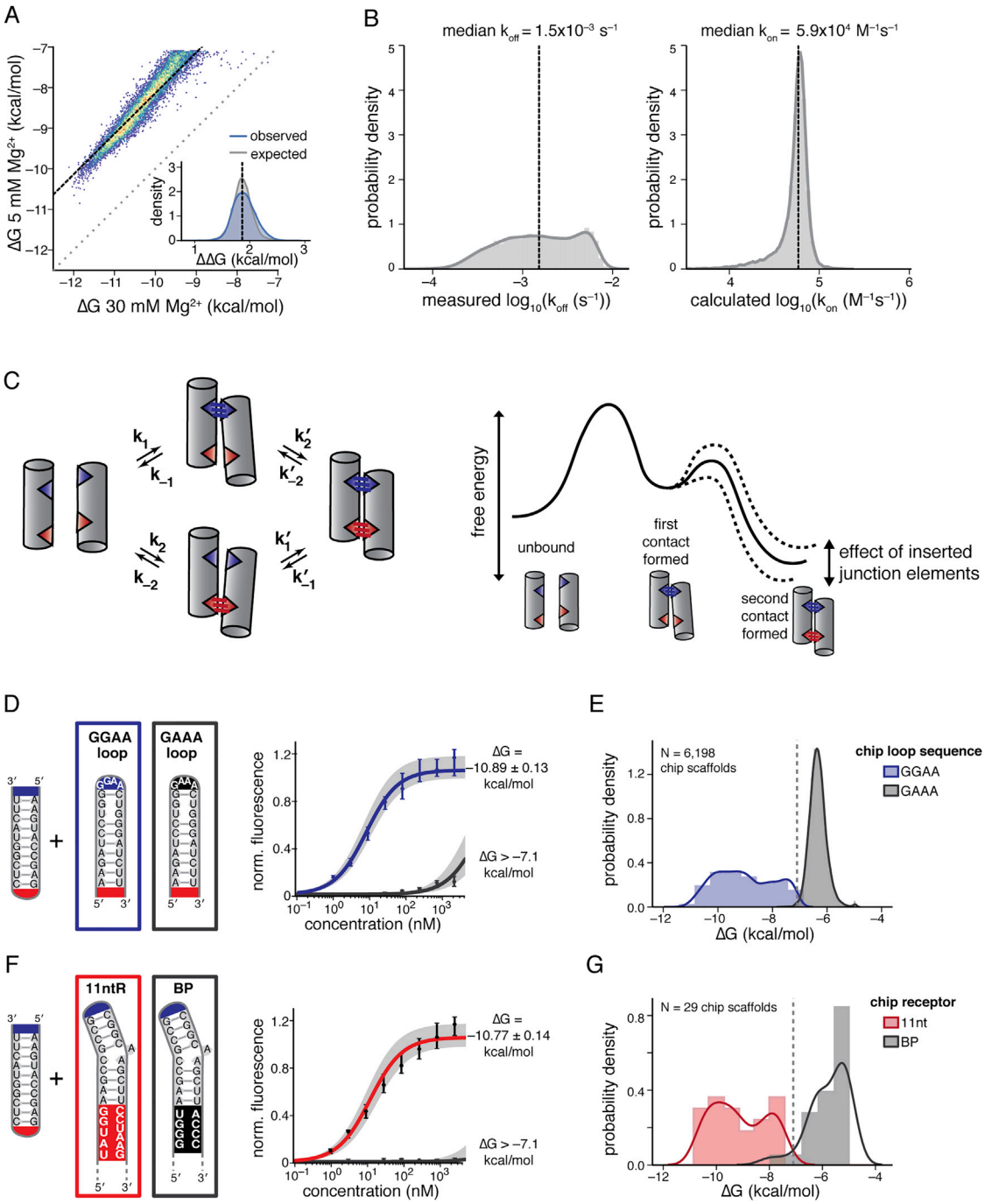


Figure S3. TectoRNA Binding Requires Formation of Both Tertiary Contacts and Is More Stable at Higher Mg^{2+} Concentration, Related to Figure 1

A) Difference in binding free energy of two Mg^{2+} conditions. Scatterplot compares the affinity of the 10-bp flow piece to each chip variant, at either 30 mM Mg^{2+} (default value) or 5 mM Mg^{2+} (89 mM Tris-Borate, pH 8.0). Inset histogram shows the distribution of ΔΔG between the two conditions across chip variants (blue), or the expected distribution given an overall, constant offset and measurement error. Similarity between observed and expected suggests a largely constant electrostatic effect from altering the magnesium concentration.

B) Histogram of measured dissociation rate constants (k_{off}) and calculated association rate constants (k_{on}) for tectoRNA variants composed of the 10-bp flow piece and ~17,000 chip piece variants with the canonical loop (GGAA) and receptor (11nt). Calculated association rate constants are a function of the dissociation rate constant and the overall equilibrium constant: $k_{on} = k_{off} / K_d$. Variants are filtered to have combined error on RTlog(k_{on}) less than 0.5 kcal/mol (95% confidence interval) and G less than –9 kcal/mol. Previous studies estimated the k_{on} as ~10^4 M^{-1} s^{-1} for a single TL/TLR motif at 125 mM Mg^{2+} (Herschlag et al., 2015; Qin et al., 2001), within an order of magnitude of the median association rate constant of 5.9 × 10^4 $M^{-1}s^{-1}$ observed here.

(legend continued on next page)

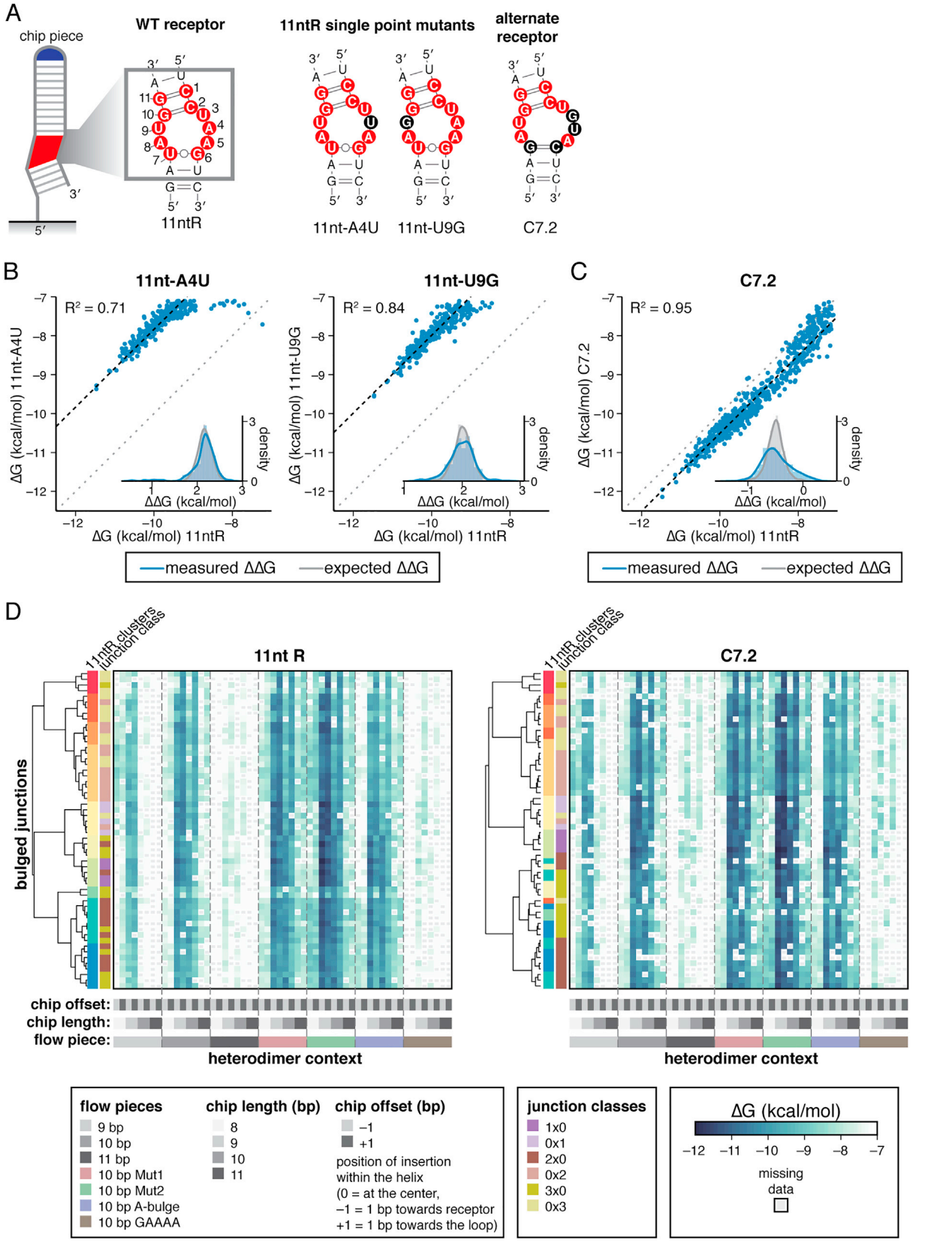
C) Kinetic model for association of tectoRNA variants. (Left) The flow and chip pieces associate through one tertiary contact and subsequently form the second tertiary contact. (Right) Free energy-reaction diagram for one pathway of the association. The constant k_{on} across inserted junction elements (E) supports that formation of the first tertiary contact is rate-limiting for all junctions. In this model, formation of the second contact is faster than dissociation of the first, regardless of the inserted junction sequence. Presumably the junctions affect the rate of forming the secondary contact once the first is formed, thus affecting the relative affinity of the complex.

D) Binding curves of a 10-bp flow piece to a chip scaffold terminated either by the wild-type loop (GGAA; blue) or mutated loop (GAAA; black).

E) Distribution of affinities (ΔG) of a 10-bp flow piece terminated by the GGAA loop or the GAAA loop binding to chip scaffolds. Dashed line indicates our applied threshold for measurable affinity.

F) Binding curves of GGAA-terminated 10-bp flow piece to a chip scaffold either with the wild-type receptor (11nt; red) or a base-paired receptor (BP; black).

G) Distribution of binding affinities for chip scaffolds, with either the 11nt or the BP receptor. Dashed line indicates our applied threshold for measurable affinity.



(legend on next page)

Figure S4. Thermodynamic Fingerprints of Junctions with Alternate Tertiary Receptors Show Consistent Results, Related to Figure 2

(A) The effect of two point mutations in the tertiary contact receptor (11nt-A4U or 11nt-U9G) and an alternate tertiary contact receptor (C7.2) were compared to the WT receptor (11ntR) across tectoRNA chip scaffolds of varying helix length and inserted junction elements.

(B and C) Scatterplots show the affinity of each point mutation (B) or the alternate receptor (C) versus the WT receptor across each chip scaffold. Black dashed line corresponds to the overall effect ($\Delta\Delta G$) between the two receptors; gray dotted line indicates $\Delta\Delta G = 0$ kcal/mol. Histograms show the distribution of the effect of the tertiary contact mutation across scaffolds: either the actual differences (blue) or expected differences (gray) assuming the constant, additive effect (shown in dashed line in scatterplot) and measurement error. N = 292, 273, or 601 scaffolds for 11nt-A4U, 11nt-U9G, or C7.2, respectively, that had measurable binding in both receptor contexts.

(D) Thermodynamic fingerprints of bulged junctions were measured either with the WT tertiary contact receptor (11ntR; left) or an alternate receptor (C7.2; right). Each set of thermodynamic fingerprints were hierarchically clustered (shown in dendrogram), and the 11ntR fingerprints were additionally assigned to classes by calling 10 flat clusters on the dendrogram (shown as colors red–blue in the colorbar labeled "11ntR clusters"). The secondary structure class is shown as the colorbar labeled "junction class." Clustering the two sets of junctions by their thermodynamic fingerprints gives largely the same results in either of the two tertiary contact receptor contexts. To quantify this observation, we asked how often junctions that clustered together with the 11ntR also clustered together in the C7.2. We found that 59% of junctions co-clustered, much greater than the 11% expected by chance, and approaching the 66% of junctions co-clustering if we recluster the 11ntR fingerprints after adding values sampled from measurement error, which serves as a ceiling for the maximum fraction of co-clustering junctions. Measurements were made for four additional flow pieces that had either different helical sequence (10-bp Mut1 and 10-bp Mut2), an additional bulged adenosine base in the helical segment (10-bp A-bulge), or an adenosine inserted in the tetraloop that interacts with the chip-piece tertiary contact receptor (10-bp GAAAA). Gray squares indicated missing measurements due to low representation or high error (> 0.5 kcal/mol).

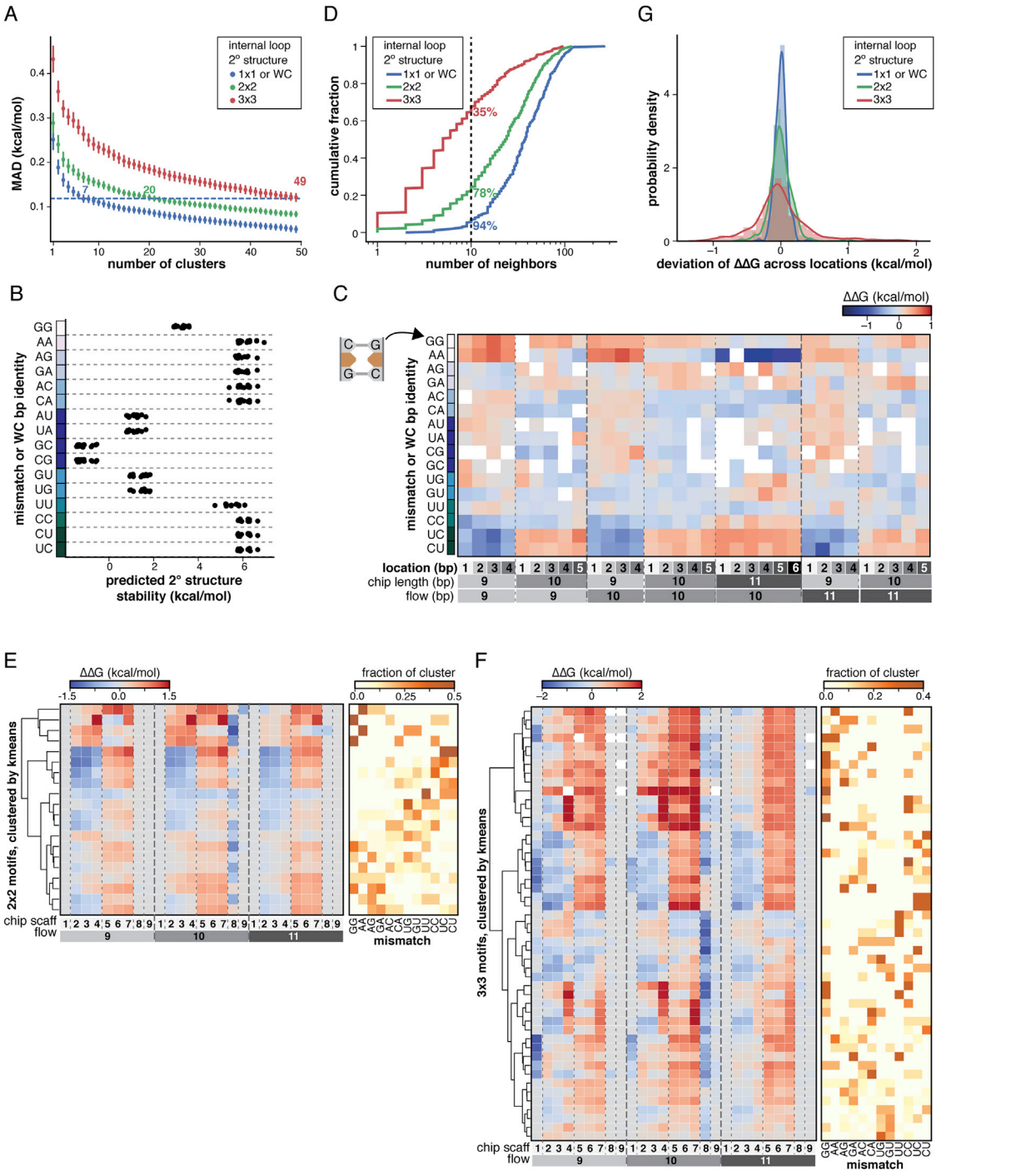


Figure S5. Affinity Fingerprints of Mismatched Motifs Revealed Strong Effect of Mismatch Identity, Related to Figure 4

(A) The average MAD between the thermodynamic fingerprint of each junction sequence and the average fingerprint of its cluster, versus the number of clusters used in k-means clustering. Horizontal line shows the average MAD for 1x1 and WC pairs using 7 clusters. (MAD expected from error is 0.1 kcal/mol.) Numbers indicate the number of clusters necessary to achieve the same average MAD for the 2x2 and 3x3 junctions. Error bars are 95% CI across junction sequences.

(B) The effect of each mismatch or bp on the secondary structure stability of a duplex, relative to the average effect across the four WC bps. Each point corresponds to the 16 different flanking bps adjacent to the mismatch/bp. Effects were calculated using RNAfold v. 2.1.8 (Lorenz et al., 2011).

(C) Binding energies for single mismatch motifs inserted at different locations in chip piece helices, where location is the number of bps between the receptor and the inserted junction element (indicated at top left). Affinity is relative to the average affinity across sequences within each chip-scaffold/flow-piece context, i.e., each column. Contexts are only shown if the majority of measurements in that context were within the measurable range of affinity (i.e., < –7.1 kcal/mol).

(legend continued on next page)

(D) Cumulative distribution plot shows the fraction of 1x1, 2x2, and 3x3 motifs that have that have less than or equal to the indicated number of "neighbors," where neighbors have similar thermodynamic behavior (see STAR Methods and Figure 6). The percentages indicate the fraction of each class of motif that has 10 or more neighbors.

(E and F) K-means clustering of thermodynamic fingerprints for (C) 2x2 and (D) 3x3 motifs (k = 20 or k = 49 clusters, respectively). Average thermodynamic fingerprints of each cluster are shown, which in turn are hierarchically clustered. Values are given relative the WC average profile, as in Figure 4C, (Right) heatmap of the fraction of each cluster with the indicated mismatch in any of the two (or three) mismatch positions.

(G) Histogram shows spread of affinity measurements across different locations for 1x1, 2x2, and 3x3 junctions. For each junction sequence, deviation of ΔG values across scaffolds 2-4 (Figure 4B; i.e., three different locations within the 9-bp chip helix), was measured with the 10-bp flow piece.

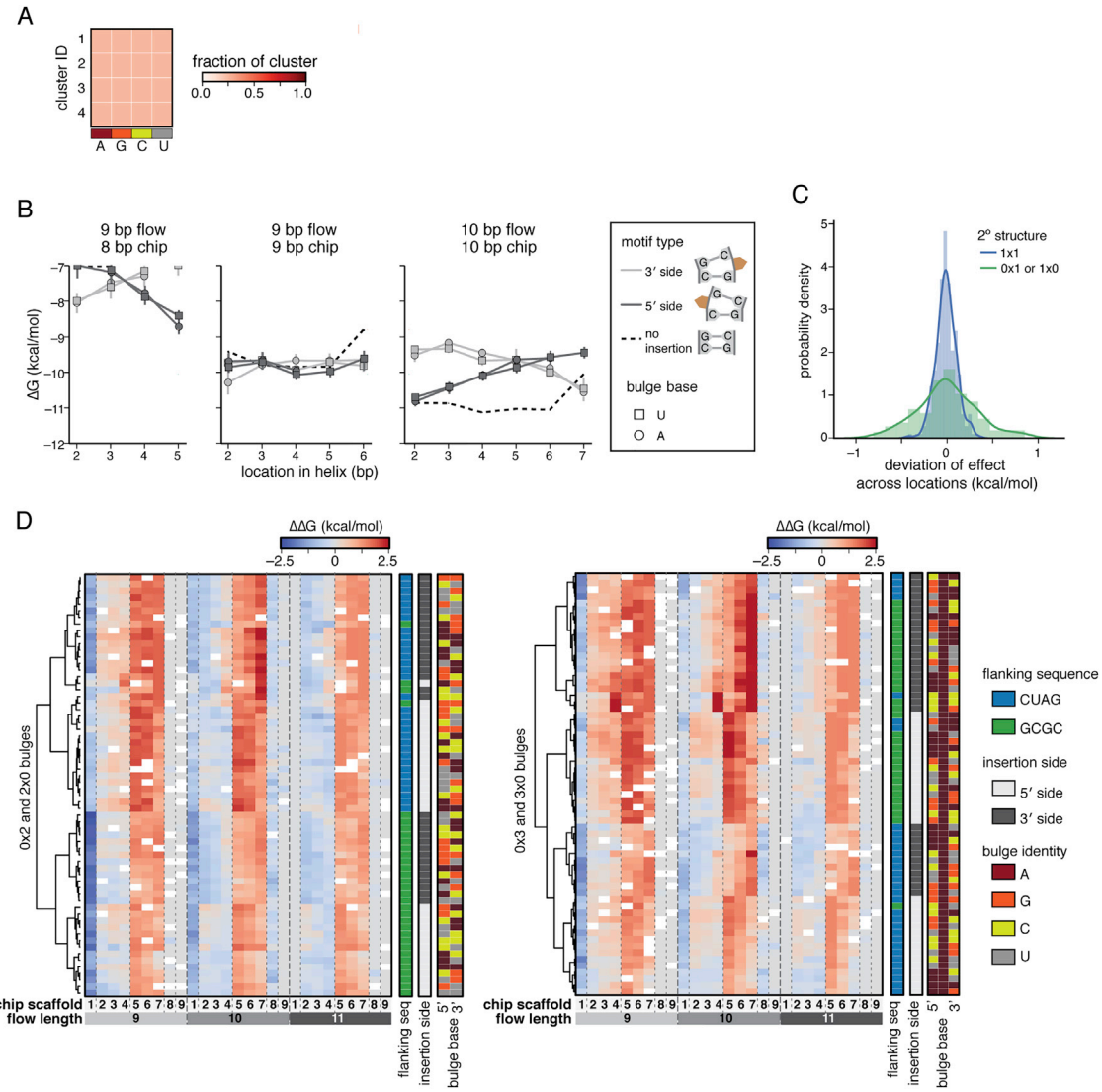


Figure S6. Effect of Bulged Residues Is Position Dependent, Related to Figure 5

(A) K-means clustering (k = 4 clusters) was carried out on thermodynamic fingerprints shown in Figure 5B. Heatmap shows the fraction of cluster members that have each of the indicated bulged residues. We observe no evident enrichment for bulge base identity in clusters.

(B) Affinity dependence on location within the helix of bulged junctions or WC pairs, where location is the number of bps between the receptor and the junction. Three different chip-scaffold/flow-piece contexts are shown.

(C) Histogram shows spread of affinity measurements across locations for 0x1 and 1x0 or for 1x1 junctions. For each junction sequence, deviation of ΔG values across locations for all flow-piece, chip-piece contexts with measureable binding.

(D) Heatmap of hierarchically clustered thermodynamic fingerprints of individual 0x2 or 2x0 bulged junctions (left) and 0x3 or 3x0 junctions (right). Three-base-bulge junctions were subsetted to always have an adenine residue in the middle bulge position.

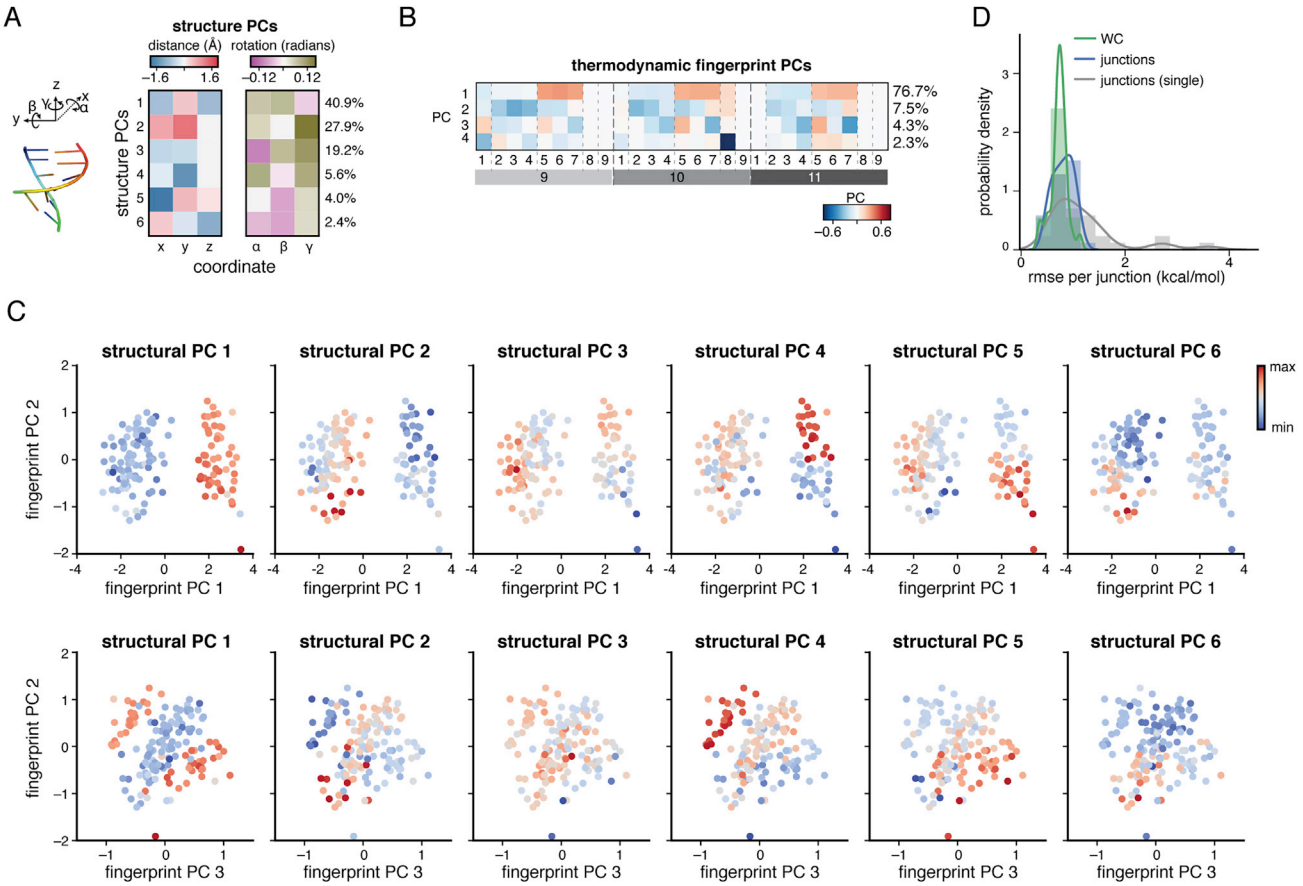


Figure S7. Specific Structural Changes Correspond to Differences among Thermodynamic Fingerprints, Related to Figure 7

(A) To compare the structural parameters of junctions, 378 structures of junctions were extracted from the PDB crystal structure database. Structures were modeled to include up to two flanking bps to replace a total of six bps in the tectoRNA chip helix. All structures were aligned by the top bp (i.e., the bp closest to the loop when inserted in the tectoRNA). This bp formed the 'origin', and a coordinate system was defined as shown, with *x*, *y*, and *z* parameterizing standard translational coordinates and α, β, γ parameterizing rotational coordinates around each of these translational axes, respectively. PC analysis was carried out (see STAR Methods); loadings for each PC of each of the six coordinates are shown, with percentages that indicate the fraction of total variance associated with each PC.

(B) To relate thermodynamic fingerprints to structural parameters, PC analysis of the fingerprints of the 148 junction sequences with 10 or more structurally characterized neighbors was first carried out. The loadings of each of the first four PCs across chip/flow contexts are shown as a heatmap. Percentages indicate the fraction of total variance along each PC.

(C) The 148 junctions are plotted by the values of their thermodynamic fingerprint PC 1 and 2 (top) or 2 and 3 (bottom). Colors indicate the value of the indicated structural PC, averaged across all structures associated with that junction sequence's neighbors.

(D) Histogram of root-mean-squared error (RMSE) between the observed and predicted affinities of each junction across each context.