Evolutionary, structural and functional explorations of non-coding RNA and protein genetic robustness

Dorien S Coray[a,b], Nellie Sibaeva[b,c], Stephanie McGimpsey[a,b,d], Paul P Gardner[a,b,d,e]

Affiliations:
a. School of Biological Sciences, University of Canterbury, Christchurch, NZ
b. Biomolecular Interaction Centre, University of Canterbury, Christchurch, NZ
c. School of Biological Sciences, University of Auckland, NZ.
d. Department of Biochemistry, University of Otago, New Zealand
e. Address correspondence to Paul Gardner (paul.gardner@otago.ac.nz).

Abbreviations:
BANP, CRISPR, Indels, ncRNA, ORF, RNase, RNP, rRNA, tRNA, RY

Classification: BIOLOGICAL SCIENCES: Biophysics and Computational Biology; Evolution; Genetics; Systems Biology.

Abstract

The reactions of functional molecules like RNAs and proteins to mutation affect both host cell viability and biomolecular evolution. These molecules are considered robust if function is maintained alongside the mutation. RNAs and proteins have different structural and functional characteristics that affect their robustness, and to date, comparisons between them have been theoretical. In this work, we tested the relative mutational robustness of RNA and protein pairs using three approaches: evolutionary, structural, and functional. We compared the nucleotide diversities of functional RNAs with those of matched proteins. Across different levels of conservation, there were no differences in nucleotide-level variations between the biomolecules. We then directly tested the robustness of the RNA and protein pairs with *in vitro* and *in silico* mutagenesis of their respective genes. The *in silico* experiments showed that RNAs and proteins reacted similarly to point mutations and insertions or deletions. *In vitro*, mutated fluorescent RNAs retained greater levels of function than the proteins, but this may be because of differences in the robustness of the specific individual molecules rather than being indicative of a larger trend. In this first experimental comparison of proteins and RNAs, we found no consistent quantitative differences in mutational robustness. Future work on potential qualitative differences and other forms of robustness will give further insight into the evolution and functionality of biomolecules.

Significance Statement

The ability of functional RNAs and proteins to maintain function despite mutations in their respective genes is known as mutational robustness. Robustness impacts how molecules maintain and change phenotypes, which has a bearing on the evolution and the origin of life as well as influences modern biotechnology. Both RNA and protein have mechanisms that allow them to absorb DNA-level changes. Proteins have a redundant genetic code and non-coding RNAs can maintain structure through flexible base-pairing possibilities. The few theoretical treatments comparing RNA and protein robustness differ in their conclusions. In this experimental comparison of RNAs and proteins, we find that RNAs and proteins achieve remarkably similar degrees of overall genetic robustness.

Introduction

Long known for its role in translation, RNA is commonly involved in controlling gene expression (e.g., bacterial small RNAs and microRNAs) (1, 2), intrinsic immunity (e.g., CRISPR-mediated acquired immunity) (3–5) and the cell's response to environmental stimuli (e.g., thermosensors and riboswitches) (6, 7). The discovery of functional RNAs, like transfer RNA and catalytic RNAs, led to the proposal of an ancestral RNA world with RNA catalyzing the reactions of life and encoding genetic information (8). Despite the importance of non-coding RNAs (ncRNAs) to cell function, many RNAs exhibit low sequence conservation and are not as broadly distributed as their proteinaceous brethren (9, 10).

The ability to preserve a phenotype in the face of sequence perturbations is termed mutational robustness (11–14). More robust molecules maintain their phenotypes despite mutations, while less robust molecules lose their function rapidly with mutation. The structural level at which mutation is considered, and the types of phenotypes that are measured, can vary between analyses. Here, we are considering mutations in DNA nucleotide sequence and how they modify phenotypes like the structure and function of encoded genes. RNAs and proteins have independent mechanisms that allow for near-neutral change within molecules, where structure and function are relatively unaffected by mutation (Table 1). RNA structure may be dominated by canonical Watson-Crick style base-pairings; however, a broad range of non-canonical interactions are possible between most base-pair combinations on each of the accessible pairing edges (15). Among these, the G:U wobble base-pair is frequently observed (22). The RNA secondary structure, and often function, can thus be preserved after mutation through a combination of canonical and non-canonical base-pair interactions (16–18).

Protein robustness relies, in part, on the robustness of the genetic code. Degeneracy of the genetic code allows for mapping of up to six codons to the same amino acid (19). Furthermore, when point mutations change the amino acid, the new amino acid coded for is likely to have similar biochemical properties due to the code's organization (20). Simulations have shown that the extant genetic code is significantly more

robust to substitution and frameshift mutations than randomly generated genetic codes (18, 21–23). Premature stop codons and frameshift errors can be introduced by substitutions or indels (insertions and deletions). While ncRNA production requires transcription—and maybe some additional maturation such as editing, splicing or cleavage—proteins require further stages of maturation such as splicing and translation, which depend upon the maintenance of a correct reading frame (24). These additional steps likely amplify the potential harm of nucleotide changes (25). This is supported by the fact that disease-associated sequence variation is enriched ten-fold in human protein-coding regions (26, 27) and that overall variation is reduced in coding regions, particularly indels (28). Because of this, we hypothesize that RNAs are more robust to mutation than proteins, and can tolerate greater sequence change while maintaining function.

Table 1: Avenues of neutral change within RNAs and proteins (29)

Mechanism	RNA	Protein
Primary sequence	Preponderance of transition mutations (R to R, Y to Y) over transversions (R to Y) maintains biochemistry (18)	Maintenance of amino acid sequence through degenerate genetic code E.g. CGN = Arg, GGN = Gly (19)
Secondary/ tertiary structure	Non-canonical base-pairing maintains stems and loops E.g. Wobble base-pairing (G:U) Covarying sites may preserve base-pairs (30) Isostericity of non-canonical base-pair interactions (31).	Point mutations may not alter the biochemistry of the amino acid, maintaining structure/function I.e. hydrophobic residues stay hydrophobic (32)
Size	Smaller molecules are more robust as they are less likely to acquire mutations at a fixed rate of mutations per kb (33, 34) .	
Stability	More stable structures may buffer mutations, as can more stable regions within the molecule I.e: Stems > loops for RNA(35) and alpha helices >beta sheets>loops for protein (36)	

Previous experimental work has focused primarily on the robustness of either RNAs or proteins (37). To date, the only direct comparisons of RNA and protein robustness have involved theoretical treatment of neutral networks with simulated datasets (38, 39). Neutral networks are all sequences (RNA or protein) that give rise to the same phenotype, and are connected by a point mutation. A mutation that keeps the sequence within the network does not affect the phenotype. Early work has suggested that RNA networks and protein networks differ both in size and shape, which affects their robustness (38–41), while more recent work suggests the two biomolecules are actually similar (39).

To begin exploring this issue, we asked whether RNAs and proteins differ overall in mutational robustness. Specifically, we investigated how mutations in DNA affect the structure and function of RNA and protein molecules using a combination of computational and experimental approaches. Given the variety of mechanisms and levels at which molecules can alter robustness (summarized in Table 1) and

the biochemical differences between RNAs and proteins, it is impossible to make fair comparisons between them. No particular molecule, with its own selection history and functional constraints, is absolutely representative of its type. Nonetheless, in exploring differences between protein and RNA robustness under comparable conditions, we may gain insight into the evolution of RNAs and proteins.

Results

We devised three independent *in silico* and *in vitro* tests to explore whether RNAs and proteins differed in their robustness to mutation. First, we considered the degree of sequence change between matched RNA and protein families. These pairs have shared functions (e.g., ribosomal RNA and protein components, or riboswitches and the proteins these generally regulate) and shared phylogenetic distributions and selection histories. Our expectation is that more robust genes will tolerate more mutations over fixed timescales and thus exhibit greater sequence change than less robust genes, allowing us to differentiate between them. Second, we mutated a functionally linked RNA and protein pair and predicted structures for these with *in silico* methods to compare their structural robustness. This is a more direct measure of robustness to mutation, but it only measures the probabilities of predicted structures and does not necessarily capture when function itself is lost. These results may be influenced more by the robustness of the computational methods than genuine mutational robustness. Finally, we compared function directly with mutagenesis of functionally comparable fluorescent RNA (42, 43) and protein (44–46), and quantified the fraction of mutants that maintained function for each class of molecule.

Sequence diversities of ncRNAs and proteins involved in the same processes

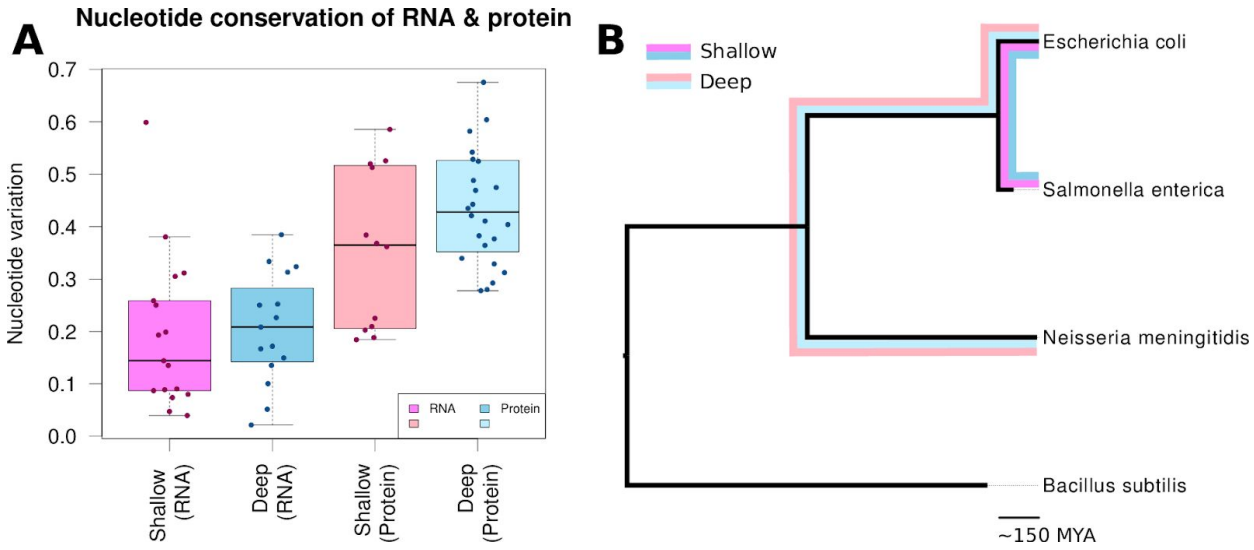

Figure 1: Proportion of variable nucleotides between shallow and deep divergence times for conserved RNA and protein families. The proportion of nucleotide variation was tabulated for aligned

orthologous RNAs (pink) and proteins (blue) from *Escherichia coli* and *Neisseria meningitidis* (deep divergence, lighter shades) or *E. coli* and *Salmonella enterica* (shallow divergence, darker shades). (**A**) Nucleotide variation is shown as the percentage of total nucleotides and indels that differ between the two species, with 50% indicating that half of the aligned positions differ. (**B**) The relationship between the deep (thin colored lines) and shallow (thick lines) diverged species on a 16S rRNA, dnaml (47), phylogenetic tree. It should be noted that the 'shallow' species diverged approximately 150 million years ago (MYA).

Our expectation is that nucleotide variation in RNAs and proteins between diverged species will indicate the degree of neutral variation that has occurred while the gene functions have been preserved. More robust genes are expected to tolerate more mutations over time. We have collected Rfam RNA and Pfam protein-domain pairs that are involved in the same biological processes (48, 49). These can be broadly classified as ribonucleotide particles (RNPs), cis-regulatory elements and their downstream protein gene pairs, and dual-function genes where one partner is modified or processed by the other (Table S1). These RNA and protein pairs have similar selection histories because of their shared functions, though their individual structural and/or catalytic constraints will vary.

We curated two lists of RNA:protein pairs and compared them across well-studied and annotated genomes. These pairs are distant enough to exhibit sequence diversity and have matched G+C contents (~50%) (Figure 1B). We selected 25 RNA and protein-domain pairs that are deeply conserved and encoded in the genomes for *Escherichia coli* (U00096.3) and *Neisseria meningitidis* (AL157959.1). These two species are from the same phylum (Proteobacteria) but hail from different taxonomic classes (Gammaproteobacteria and Betaproteobacteria). The selected genes or domains are primarily core genes (e.g., ribosomal RNA and ribosomal proteins, tRNA and tRNA synthetases (Table S1)) involved in transcription or translation and are deeply conserved. We also collected 18 less-conserved 'shallow' pairs encoded in the genomes of *E. coli* and *Salmonella enterica* (AE014613.1), which are from the same family (Enterobacteriaceae) but different genera (Escherichia and Salmonella). These RNA and protein-domain pairs share a recent evolutionary history—Escherichia and Salmonella diverged approximately 150 million years ago(50) and represent families with limited phylogenetic distributions and possible tolerance of a high mutation rates because of a relaxed level of selection that is characteristic of new genes (51, 52).

At each level of conservation, we computed the number of nucleotides that varied between the two species for a given gene. After normalizing by the length of the nucleotide sequence, we obtained a measure of the nucleotide variation for each gene. We then compared the sequence diversities of the conserved RNAs with those of matched protein-coding genes or domains (Figure 1). Each divergent nucleotide (mutation) was further classified as neutral or not, depending on whether it preserved secondary structure (RNA), amino acids (protein) and/or the biochemistry of the RNA or protein (Figure S1).

For both the shallow and deep groups, the number of mutations per nucleotide did not differ significantly between the proteins and RNAs ($p = 0.4493$ and 0.06965, respectively: Wilcoxon rank-sum test) (Figure 1). It is possible that interactions between the RNAs and proteins constrained the degree of variation between the two, with one of the pair evolving slower because it maintained interactions with a slowly evolving partner (53). We did not, however, see a correlation between the rate of nucleotide variation in a

given RNA and its matched protein (Figure S2), leading us to conclude that this was most likely not the case.

The sequence diversity between populations provides an indication that a gene's function is robust to changes in the nucleotide sequence. By this measure, RNAs and proteins exhibit similar degrees of robustness.

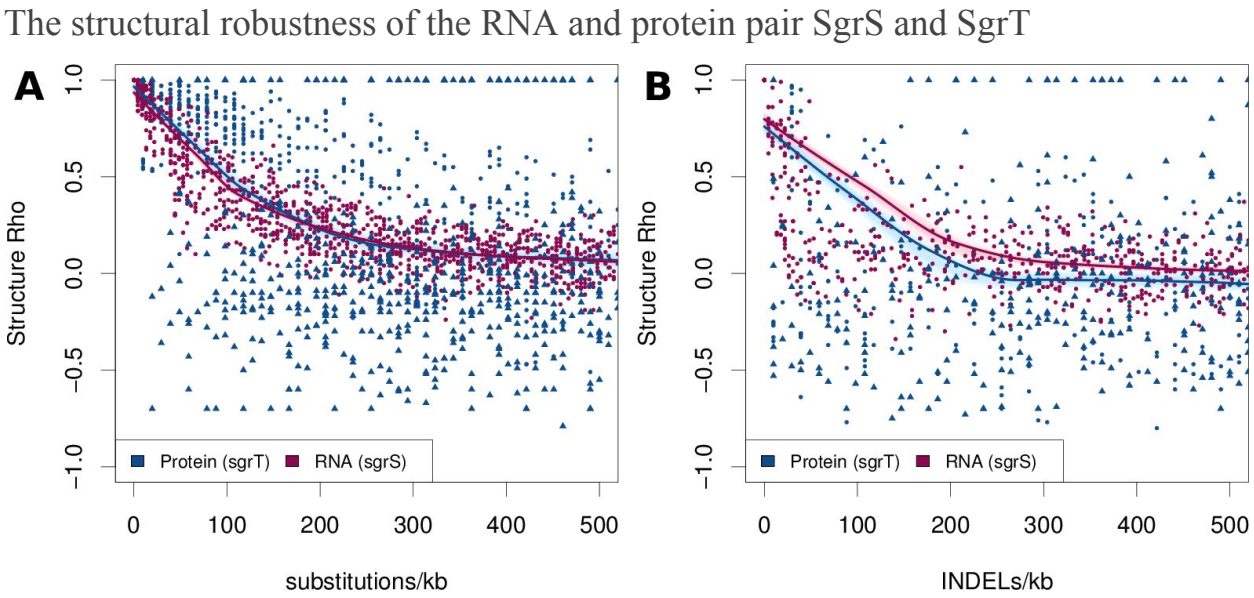

Figure 2: Robustness of structure predictions to random *in silico* mutagenesis for a protein (SgrT) and non-coding RNA (SgrS). Random mutants of the SgrT messenger RNA (blue) and the SgrS small RNA (pink) were generated *in silico*. Secondary structure probabilities for each were predicted using "PSSpred" and "RNAfold-p". The per-residue probabilities of alpha/beta/coil structures (protein) or base-paired/not-base-paired (RNA) were compared between native and mutated sequences using Spearman's correlation. This gives a "structure rho", where 1 implies the predicted mutant structure is identical to the predicted parental structure, 0 means there is no correlation between the two, and −1 corresponds to a perfect inverse correlation. (**A**) Substitution mutations and (**B**) insertion or deletion mutations (indels) were introduced into the protein (blue) and RNA (pink) at rates ranging from 1 to 500 mutations per kilobase (kb). Points corresponding to truncated protein or sRNA with a length less than 75% that of the original are indicated with a solid triangle, otherwise a solid circle is used. The average trends between mutation rates and structure rho are indicated with local polynomial regression (loess) curves. To indicate the confidence for each loess curve, these were bootstrapped 500 times and plotted in light pink or blue to resampled points.

The nucleotide sequences of a representative *E. coli* ncRNA and protein were mutated *in silico*. BThe protein sequence was then translated, and the secondary structures of the ncRNA and protein mutants were predicted using *in silico* methods and compared with the predicted parent structures. This provided a measure of how robust the structure was to mutation. We selected a

ncRNA:protein pair that met the following criteria: 1. Both the RNA and protein were structured, that is, the tertiary structure was important for function. 2. The RNA and protein were short as the computational requirements scale poorly with sequence length. 3. The protein structure was not contained in the Protein Data Bank (PDB) snapshot used by the protein secondary structure prediction (PSSpred) tool. The SgrS RNA (Rfam accession: RF00534, 227 nucleotides long) and corresponding SgrT protein (Pfam accession: PF15894, 102 nucleotides long) pair met these requirements. SgrS is an Hfq-binding, antisense regulatory RNA that encodes a short peptide, SgrT(54). SgrS and SgrT act synergistically during periods of glucose-phosphate stress: the RNA binds to a number of messenger RNAs (mRNA) while the protein acts as a regulator (55).

The per-residue secondary structure probabilities from "RNAfold-p" and "PSSpred" mutant sequences were compared with structure probabilities for the parental sequences. This gives a "structure rho" (Spearman's correlation coefficient) score. SgrT and SgrS respond similarly to point mutations and indels. Both the RNA and protein mutants retained about half the parental structure (structure correlation of 0.5) on average at 100 point mutations per kilobase, though the protein showed a more variable response to mutation than RNA (Figure 2). The ability of folded proteins to undergo structure flips from predominantly helical to alternative conformations results in a greater number of negative correlations. There were a small number of indels that had a stronger effect than point mutations, but the correlations of RNA and protein still reached 0.5 with approximately 100 indels per kilobase. The protein was slightly more sensitive to indels than RNA, but showed a similar overall level of decline in its structure. The truncated proteins (triangles, Figure 2), are the result of premature stop codons. These cause small sample size effects resulting in more extreme correlations (both high and low).

The structure analysis may be influenced by differences in the RNA and protein structure prediction methods. The protein structure inference uses automatically generated sequence alignments with a snapshot of the National Center for Biotechnology Information's (NCBI) non-redundant (NR) protein sequence database, plus a machine-learning method to estimate the probabilities of different secondary structure elements(56). By contrast, the RNA structure inference is solely based on the sequence, and uses a nearest-neighbour energy model (57, 58). RNA folding is notoriously difficult as small parameter changes (in the energy model or sequence) can result in very different minimum free energy structure predictions. However, our approach, which is based on the Boltzmann distribution, can somewhat mitigate this issue (59) by selecting a protein with few homologs in the NR and PDB databases. We ran the same analysis with the CsrA/CsrB protein:RNA pair (Figure S3), which do have protein homologs in the databases. As expected this showed an increase in protein robustness, which was likely due to matches with CsrA homologs in the NR database and the solved CsrA structure in the PDB, in spite of the random mutations we introduced.

This result is a measure of predicted structural robustness. It is possible that the structure could be maintained but function lost, or that some molecules may continue to function better than others despite changes in structure (i.e., they are more robust). Therefore, we also tested for robustness of function. To do this, we mutated an RNA and a protein matched for an assayable function (fluorescence) and tested these mutations *in vivo*.

Mutational robustness of a functionally equivalent RNA and protein

A

	Insertion	Deletion	Transitions	Transversion
mCherry (protein)	**4**	**0**	**20**	**18**
Broccoli (RNA)	**5**	**3**	**19**	**18**

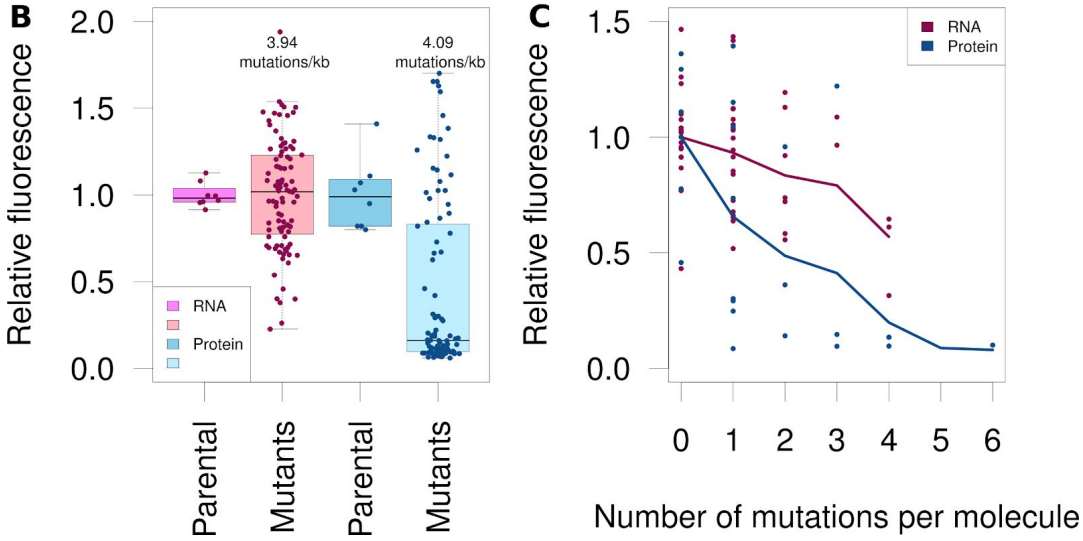

Figure 3: Relative fluorescence intensities of mutated RNA Broccoli and mutated protein mCherry.

We generated libraries of the randomly mutated fluorescent RNA aptamer Broccoli and fluorescent protein mCherry, which we then tested for function relative to an unmutated control. (A) The mCherry and Broccoli libraries were matched for similar rates of mutations per kilobase (kb) (4.09 and 3.94, respectively) using an error-prone PCR protocol. The fluorescence intensities for 96 mutants each of the RNA and protein were compared with those for eight unmutated controls. Measurements were recorded for three separate replicates. (B) Individual molecules of mCherry and Broccoli mutants were sequenced and their fluorescence compared using the number of mutations per molecule (zero to six). (C) We counted the different types of variants that were observed in the sequenced mutants.

To investigate how biomolecules may differ in their robustness to mutations in DNA, we constructed mutant libraries of the fluorescent RNA aptamer Broccoli (42, 43) and the fluorescent protein mCherry (46). Both these molecules have been developed synthetically in the laboratory, and have not been subjected to strong evolutionary pressure outside of fluorescence. With a mutation frequency of approximately four mutations per kilobase, the relative fluorescence intensity for the population of Broccoli mutants was significantly ($P = 1.5 \times 10^{-13}$, Wilcoxon rank-sum test) more than that for mCherry (Figure 1A). Though the median fluorescence of the Broccoli population decreased slightly as the frequency of mutations increased, even at six mutations per kilobase, the Broccoli library had higher relative fluorescence intensities than the mCherry library with four mutations per kilobase (Figure S3). At 234 bases, the gene for Broccoli is much shorter than that for mCherry (711 bp). We sequenced approximately 40 molecules from each library and compared the number of mutations per molecule. Broccoli retained more of its fluorescence than mCherry with the same amount of mutations per molecule (Figure 1B). The frequencies of different types of mutations that occurred in the biomolecules were roughly similar, with few insertions or deletions (indels) and similar numbers of transitions and transversions.

Discussion

We hypothesized that RNA would be more robust to mutation than proteins. This is supported by the fact that RNA often requires less processing than protein to produce functional molecules (60), is not susceptible to frameshift mutations (61) and is less likely to be found over broad evolutionary distances with homology searches, possibly because of generally higher mutation rates (10). Our multi-scale tests of RNA and protein robustness revealed no consistent evidence that RNAs are more robust than proteins.

If RNAs were more robust than proteins, we would expect phylogenetically and functionally matched RNA families to have more nucleotide diversity than proteins of the same evolutionary background. By comparing the diversity of extant RNAs and proteins, we found that they actually had similar rates of nucleotide substitution. This was true across both highly conserved and more recently acquired interacting protein and RNA families. Individual RNAs and proteins tested directly for robustness to mutation through mutagenesis gave variable results. When the fluorescent protein mCherry and fluorescent RNA Broccoli were tested *in vivo*, the protein was significantly less robust than RNA to substitution mutations. By contrast, the predicted structural probabilities of the matched RNA and protein SgrS and SgrT showed that both substitution and insertion or deletion mutations degraded the structures at very similar rates.

The genetic robustness of proteins is unexpected for a few reasons. Proteins differ from RNAs in that they must be translated as well as transcribed, potentially making them more sensitive to mutations, particularly frameshifts. Robustness against translational errors reduces the chance of creating misfolded proteins but adds additional constraints to the nucleotide sequence and actually reduces the nucleotide diversity in highly expressed, highly translated proteins in yeast

and bacteria (62). We then speculated that proteins would have less sequence diversity than RNAs, which are not translated at all. Instead, we found that the proteins had at least as much sequence diversity as matching RNAs. Additionally, indel mutations can cause frameshifts in proteins, changing all downstream amino acids. RNA, having no code to protect, could potentially absorb additional nucleotides in bulges, leaving nucleotides out of stems (35, 63, 64) without dramatically affecting other regions of the RNA structure. Nonetheless, the predicted structures for RNAs were as sensitive to indels as the proteins. The extant genetic code is more robust to frameshifts than randomly generated genetic codes (23), and it seems to confer more robustness than plasticity within RNA.

One test that stood out from the others was our functional test with the fluorescent RNA aptamer Broccoli and fluorescent protein mCherry. Here, the Broccoli RNA was more robust than the mCherry protein. We used a double Broccoli aptamer, which means a mutation in one half could leave the other half with some functionality. Otherwise, the difference could be caused by the individual natures of the molecules tested. Robustness varies significantly between individuals because of many factors, including the stability of the molecule (33, 36, 65), the need to preserve interactions with other molecules and the expression level (62, 66). Both entire molecules and regions within molecules benefit from stability as it provides some buffer for destabilization caused by a mutation (36, 67) and helps the molecule maintain its structure. Using thermodynamic models of secondary structure in RNA and experiments with small RNA viroids, stems were found to be more robust than loops and to stabilize the structure of the molecule as a whole (37, 68, 69). In proteins, alpha helices were more robust than beta strands and both were more robust than unstructured coils, primarily because of the higher number of residue interactions in helices than strands or coils (33, 36).

Our investigation of the comparative robustness of RNAs and proteins was, in part, initiated by the observation that RNAs and proteins are differentially distributed across phylogenetic distances (10). We proposed that this might be due to rapid nucleotide divergence of functional RNAs making them difficult to detect, but we did not observe any significant difference in the level of nucleotide variation between RNAs and proteins at matched evolutionary divergences. Furthermore, we found no convincing evidence that RNAs are more robust to mutagenesis as a whole. If RNAs are not more robust than proteins, as our experiments imply, the more likely explanation for differences in phylogenetic distributions is that a protein homology search is statistically more powerful than that for nucleotides (70, 71) and that gene turnover and neofunctionalization are more rapid for RNAs than for proteins.

This is the first experimentally driven comparison of the robustness of RNAs and proteins. Previous comparisons have involved computational analysis of neutral networks: a collection of related sequences that give rise to the same phenotype. Earlier analysis using reduced genetic codes (e.g., G+C for RNA and hydrophobic:hydrophilic for protein) (38–41), found that RNA networks differed quantitatively and qualitatively from protein networks. More recently, Greenbury et al. found this to be dependent on the mathematical framework used, and suggested that RNAs and proteins were more similar (39, 41). Our work corroborates this, suggesting that

RNAs and proteins have similar overall robustness to mutation. This does not mean that RNAs and proteins do not differ in their responses to mutation, but that these differences at least tend to even out. The majority of RNA mutations preserve base-pairing relationships, while the majority of protein mutations preserve the biochemistry of the coded amino acid structure. Both of these ultimately preserve the molecule's structure (Figure S2). It has been proposed that protein networks are disconnected, with each phenotype separated from other phenotypes by a span of non-functional space. By contrast, RNA is more interconnected, with phenotypes close to each other in sequence space(38, 41). Such differences in neutral networks could mean that RNAs could both lose and gain functions more easily than proteins (38, 72)

Future theoretical and experimental treatments are needed to explore the initial results presented here. Acknowledging the limitations of using just one RNA and protein pair for analysis, mutagenesis could be repeated with different RNAs (e.g., Spinach (73), iSpinach (74), Mango (75)) and phylogenetically distinct proteins (e.g., GFP, luciferase, ZsGreen1, ZsYellow1 (76)). Additional comparisons of *in silico* structural robustness could utilize Boltzmann structure ensembles for structural comparisons and methods similar to FATHMM (Functional Analysis through Hidden Markhov Models) for estimations of neutral mutations.

Simulated evolution experiments could be performed to identify differences in RNA and protein evolvability as well as mutational robustness. For example, starting from random pools of sequences, RNAs and proteins could be compared directly using systematic evolution of ligands by exponential enrichment (SELEX) for RNAs (77) and directed evolution for proteins (78, 79). This approach may also be modeled computationally using methods akin to genetic algorithms, such as the flow reactor, which iteratively optimizes a random pool of sequences to fold into predetermined predicted structures (80, 81). Forms of robustness other than mutational robustness, like the robustness of RNA and protein interaction networks, could also be explored.

Robustness interplays strongly with evolvability (39, 67, 82, 83), and work on this topic can inform our understanding of how new functions evolve in proteins and RNAs going back to the evolution of life itself. How did early biomolecules function in what is likely to have been a high mutation environment (34), and would robustness to mutation have affected the transition from an RNA world to a protein world? Robustness can also help us look into the future, as we engineer smart biomolecules capable of functioning within the host cell despite inevitable mutations. While RNA and protein may not differ quantifiably in robustness, such qualifiable differences will further our understanding of the rise and fall of new functions and families of self-replicating biomolecules.

Methods

We devised three tests to explore the relative robustness of ncRNA and protein.
First, we considered the degree of sequence change between matched RNAs and proteins that were either shallow (recently diverged) or deep (conserved) (Natural variation of ncRNA:protein systems), and classified these changes as either neutral or not (Classification of near-neutral variation). Second, we tested robustness to mutation directly by mutating a specific protein and RNA pair and testing their structural robustness *in silico* (Simulated variation and ribonucleoprotein secondary structure). Finally, we tested an ncRNA and protein pair and tested their functional robustness *in vivo* (Fluorescent protein and RNA comparison).

Natural variation of functional ncRNA:protein systems

We selected pairs of RNA and protein families from Rfam and Pfam that were linked by either direct interactions or by process, and conserved over either deep phylogenetic distances (between *E. coli* and *N. meningitidis*) or shallow phylogenetic distances (between *E. coli* and *S. enterica*). The pairs included directly-interacting RNAs and proteins (components of the ribosome, RNAse P and SRP); cis-regulatory elements and their downstream genes (e.g., the cobalamin riboswitch and the TonB-dependent receptor involved in cobalamin uptake); and dual-function genes that encode both proteins and structured ncRNAs (e.g., the tryptophan operon leader and SgrS). The full list of partners is detailed in Table S1. Each pair of deep or shallow diverged nucleotide sequences was aligned, either using a Rfam covariance mode (48) and cmalign (84) or, for the protein domains, using hmmalign (v3.1b2) (85) and concordant codon-aware nucleotide alignments generated with PAL2NAL (86). The number of variant sites was recorded for each alignment and each variant was classified as either neutral or not, based upon a number of structural and biochemical models (see Figures 1 and 2).

Classification of near-neutral variation

For RNA, we considered two models of near-neutral mutation. The first, which we describe as biochemically “neutral”, incorporated transition mutations. That is, purines replacing purines and pyrimidines replacing pyrimidines (A<->G and U<->C). The second model took into account the consensus RNA secondary structure, defining as neutral mutations in loop regions and those that are structurally neutral, either compensated for by a covarying site or by a wobble mutation (e.g., A:U <-> G:U <-> G:C). For proteins, we considered three models of near-neutral mutation. The first, which we describe as “degeneracy”, identified mutations that did not change the genetic code (i.e., synonymous mutations) and those that changed the amino acid sequences (i.e., non-synonymous mutations). The second model (BANP) coded amino acid sequences into a four-letter alphabet that split the amino acids into the following broad categories: basic (B = H or

R), acidic (A = D or E), non-polar (N = A, F, L, I, M, P, V or W), and polar (P = C, N, Q, S, T or Y). Nucleotide variations that did not change the BANP sequences were considered neutral, and changes that altered it were considered non-neutral. Finally, the "blosum" model considered each nucleotide change. Those changes that resulted in an amino acid replacement and had a positive score in the BLOSUM62 matrix (32) were considered neutral. By contrast, those replacements that had a score of zero or less were considered non-neutral.

Simulated variation and ribonucleoprotein secondary structure

We mutated *E. coli* CsrB RNA and CsrA protein *in silco* and computed the abilities of these mutants to maintain their secondary structures. Both molecules are comparatively short (360 nucleotides and 52 amino acids for CsrB and CsrA, respectively), which made the following computation possible. For CsrA and CsrB, the DNA sequences were replicated with random mutations 100 times using a Perl (v5.26) script (structureMutagenerator.pl). The mutations were either substitutions or indels and the mutation rate was between 0 and 500 per kilobase. To predict the secondary structures of these mutants, we used PSSpred to infer the probability of each residue in a protein sequence forming an alpha helix, beta sheet or coil (87, 88). PSSpred uses multiple sequence alignments from PSI-BLAST searches of the NCBI NR database. The resulting alignments are fed to a combination of seven neural network predictors, which are trained to infer structures from profiles. For the RNA sequences, we used "RNAfold-p", an implementation of McCaskill's RNA partition folding function (58) found in the Vienna RNA Package (89).

Fluorescent protein and RNA Plasmid and Library construction

The fluorescent protein vector was constructed by inserting the mCherry gene into the NcoI and PmeI sites of pBAD-TOPO/LacZ/V5-His (Invitrogen) deriving pMCH01 (P_{BAD}-mCherry, pBR322+ROP backbone, Amp^R). Plasmid pBRC01 (T7-Broccoli-Broccoli, pBR322+ROP backbone, Kan^R) was purchased as pET28c-F30-2xdBroccoli (Addgene). Mutagenesis libraries were constructed using GeneMorph II Random Mutagenesis Kit (Agilent Technologies). The mCherry gene and Broccoli aptamer were amplified from their respective plasmids using Mutazyme II DNA polymerase to generate mega primers for MEGAWHOP whole plasmid PCR (90). Parental plasmids were digested with restriction enzyme DpnI, and the resulting mutation library was introduced into competent *E. coli* BL21(DE3) (Broccoli) or *E. coli* BL21(DE3) pLys (mCherry). We constructed two mCherry libraries with mutation rates of approximately one and four mutations per kilobase, and three Broccoli libraries with mutation rates of four, five and six mutation per kilobase. Approximately 10 clones from each library were sequenced to determine the mutation frequencies and whether the mutations were indels, transitions or transversions. Individual clones (n = 96) from each library were frozen for later analyses.

Fluorescent protein and RNA Fluorescence measurements

Cultures were grown at 37°C in Luria Bertani broth supplemented with appropriate antibiotics in a dry shaking incubator at 150 rpm. Each library was grown overnight in a 96-well plate before transfer to a second plate containing fresh medium supplemented with 1 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) and 200 μM DHFB-T1 (Lucerna) to induce expression of Broccoli or 0.2% arabinose to induce expression of mCherry. We also prepared a plate containing eight wells of induced parental constructs (positive), uninduced parental constructs (negative), and LB supplemented with inducers (blank) for controls. The next morning, each library plate was used to culture three independent replica plates (three total cultures per mutant) and the control plate was used to culture one replica plate (eight total cultures per control condition). All plates were grown for 6 h before a Fluostar Omega plate reader (BMG Labtech) was used to measure the optical density (600 nm) and fluorescence. Fluorescence for the mCherry mutant library was measured with a 584 nm excitation filter and a 620 nm emission filter, with a 1500 gain. Fluorescence for the Broccoli mutant library was measured with a 485 nm excitation filter and a 520 nm emission filter, with a 1000 gain. Relative fluorescent units (RFU) was divided by optical density to derive a "Growth modified RFU", and then by no-mutant controls to get the "Relative Fluorescence". The no-mutant controls for the libraries were the parental plasmids and the no-mutant controls for the individual clones were unmutated clones within the library.

Data and software availability

All the software, documentation, sequences, and results for this project are available on our github repository: https://github.com/Gardner-BinfLab/robustness-RNP.

Furthermore, the datasets used to generate Figures 1–3 and S2 are available in the following Google Sheet: https://docs.google.com/spreadsheets/d/1exZaYpTQRfTpdNBVaIOJID3Uzw_WIJy0XBOTmPaOSi4/edit?usp=sharing

Acknowledgements

We thank Gabrielle David, PhD, for editing a draft of this manuscript.

Supplementary Material

Table S1: Conserved and Recent interacting RNAs and proteins

Process	RNA	Protein domains
Deep diverged genes (*E. coli* to *N. meningitidis*)		
Ribosome function	5S rRNA (RF00001)	L5 protein (PF00281)
	SSU rRNA (RF00177)	Ribosomal proteins S21, S15 & S17 (PF01165, PF00312, PF00366)
	LSU rRNA (RF02541)	Ribosomal proteins L6, L22 & L31 (PF00347, PF17136, PF01197)
Amino acid transfer	Arginine tRNA (RF00005*)	Arginine tRNA synthetase (PF00750, PF03483)
	Leucine tRNA (RF00005*)	Leucine tRNA synthetase (PF00133, PF13603)
	Phenylalanine tRNA (RF00005*)	Phenylalanine tRNA synthetase (PF01588, PF03483)
	Serine tRNA (RF01852)	Serine tRNA synthetase (PF02403, PF00587)
tRNA processing	RNase P RNA (RF00010)	RNase P protein (PF00825)
RNA polymerase regulation	6S RNA (RF00013)	Sigma 70 (PF04546)
Ribosome rescue	tmRNA (RF00023)	Elongation factor Tu (PF00009), Ribosomal protein S1 (PF00575)
Protein trafficking	SRP RNA (RF00169)	SRP protein (PF02881, PF00448, PF02978)
Ribosome regulation	S15 leader (RF00114)	Ribosomal protein S15 (PF00312)
Thiamine pyrophosphate metabolism	TPP riboswitch (RF00059)	Phosphomethylpyrimidine synthase (PF13667, PF01964)
Shallow diverged genes (*E. coli* to *S. enterica*)		
Cis-regulatory element and downstream related genes	Cobalamin (Vitamin B12) riboswitch (RF00174)	TonB dependent receptor domains (PF00593, PF07715)
	Flavin mononucleotide riboswitch (RF00050)	DHBP synthase (PF00926)
Cis-regulatory thermoregulator and downstream gene	Repression of heat shock gene expression element (ROSE_2, RF01832)	Heat shock protein HSP20 (PF00011)
	CspA, cold shock regulator (RF01766)	Cold shock DNA binding domain (PF00313)
Dual-function structural RNA and encoded peptide regulating amino acid biosynthesis and magnesium transport	Histidine operon leader (RF00514)	Histidine leader peptide (PF08047)
	Leucine operon leader (RF00512)	Leucine leader peptide (PF08054)
	Threonine operon leader (RF00506)	Threonine leader peptide (PF08254)

	Tryptophan operon leader (RF00513)	Tryptophan leader peptide (PF08255)
	Magnesium sensor (RF01056)	Magnesium leader peptide (PF17059)
Dual functioning small RNA and encoded polypeptide regulating sugar transport	Sugar transport related sRNA SgrS (RF00534)	SgrT (PF15894)
RNP of small RNAs and protein that facilitates target binding	GcvB (RF00022), RseX (RF01401), OxyS (RF00035), MicA (RF00078)	Hfq (PF17209)
RNPs involved in glucosamin biosynthesis	GlmY RNA activator (RF00128)	RNase adapter protein RapZ domain (PF03668)
	GlmZ RNA activator (RF00083)	RNase E (PF10150)
RNP that regulates glycogen biosynthesis	CsrB (RF00018)	CsrA (PF02599)

A

Serine tRNA, loop D

```
E.coli      UACCGGGGGUUCAAAUCCCCC
N.meni      UC-CGUGAGUUCGAAUCUCAC
SS_struc    ...,<<<<<_______>>>>>
SS          .YY..Y.Y....Y....Y.Y.
RY          .NN..N.Y....Y....Y.N.
```

Length: 21
 # mutations: 7
SS: 7/7
RY: 3/7

B

Seryl-tRNA synthetase

```
E.coli(nuc) ------GAAGATATCGAGCCT
E.coli(aa)  .  .  E  D  I  E  P
N.meni(nuc) AAACATGAAGAGGCGCAGGTG
N.meni(aa)  K  H  E  E  A  Q  V
Degeneracy  NNNNNN.....NNNNN..NNN
BANP        NNNNNN.....YYYYN..YYY
BLOSUM      NNNNNN.....YNNNY..NNN
```

Length: 21
 # mutations: 14
Coding: 0/14
BANP: 7/14
BLOSUM: 2/14

C

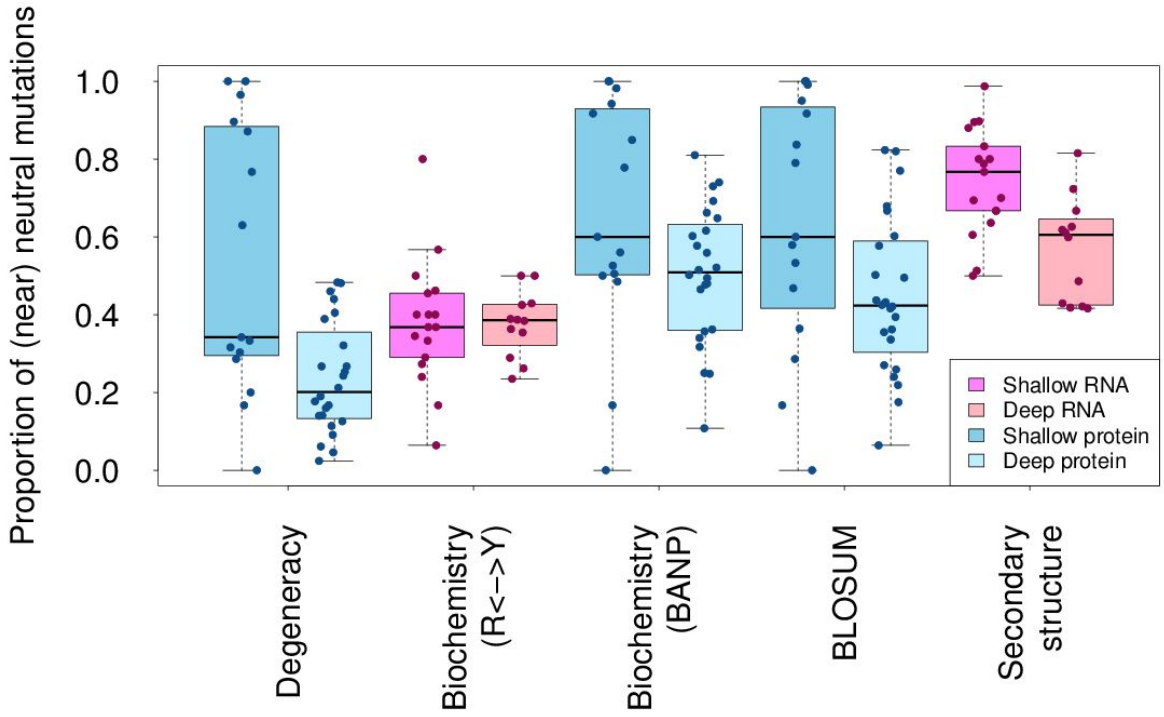

Figure S1: The proportion of nucleotide variants in RNA or protein that can be classed as neutral. A collection of functionally linked RNA and protein families that are shared between *E. coli* and *N. meningitidis* (*N. meni*) (Deep, lighter shades) or between *E. coli* and *S. enterica* (Shallow, darker shades). Each nucleotide variant is classified as either neutral or non-neutral according to a number of different models. (**A&B**) Exemplar genome variants and different classification schemes. (**A**) To score differences in the RNA serine tRNA, for example, secondary structure of each was determined (**SS_struct**) and changes between species (in red) was scored as either near-neutral or not, for changes in secondary structure (**SS** or **Secondary structure**) or biochemistry (**RY** show transitions, R: A<->G, Y: C<->U). (**B**) To score differences in the protein seryl-tRNA synthetase. For example, both nucleotide (nuc) and amino acid (aa) sequences were compared across the two species. The nucleotide differences between species was scored as neutral if the resulting amino acids were the same, labelled **Degeneracy**. Biochemically

neutral variation, labelled **BANP,** classed the following groups of replacements as neutral (**B**asic (H,R,K), **A**cidic (D,E), **N**on-polar (F,L,W,P,I,M,V,A) or **P**olar (G,S,Y,C,T,N,Q)) or if amino acids replacements were assigned a non-negative score in the **BLOSUM** score matrix (32). (**C**) The proportion of near-neutral mutations for each RNA or protein was compared for different models of neutrality across deep and shallow phylogenetic distances for RNAs and proteins. The x-axis labels are described above.

Presuming that molecules surveyed were still functional in both species, these results tabulate all variations that preserved function of the molecules In both conserved and less conserved RNAs, of all the functional mutations, secondary structure was preserved more than biochemistry (Figure S2B). In proteins, biochemistry of the coded amino acid was preserved more than other traits (Figure S2B). This may reflect that, as functional molecules diverge, changes in one part of the molecule allow for compensatory change in other regions, which increases the overall diversity. Each divergent nucleotide (mutation) was further scored as presumed neutral or not, depending on whether it preserved secondary structure (RNA), amino acids (protein) and/or biochemistry (RNA or protein) (Figure S2A). Near-neutral protein biochemistry was scored using two metrics. BLocks SUbstitution Matrix (BLOSUM) is based on the frequency of one amino acid substitution for another in related proteins. A near-neutral BLOSUM mutation would be one that has a positive score; thus, is a common substitution during evolution (32). BANP categorizes all amino acids as either basic, acidic, nonpolar or polar. A near-neutral BANP mutation would be one where the amino acid stays in the same category.

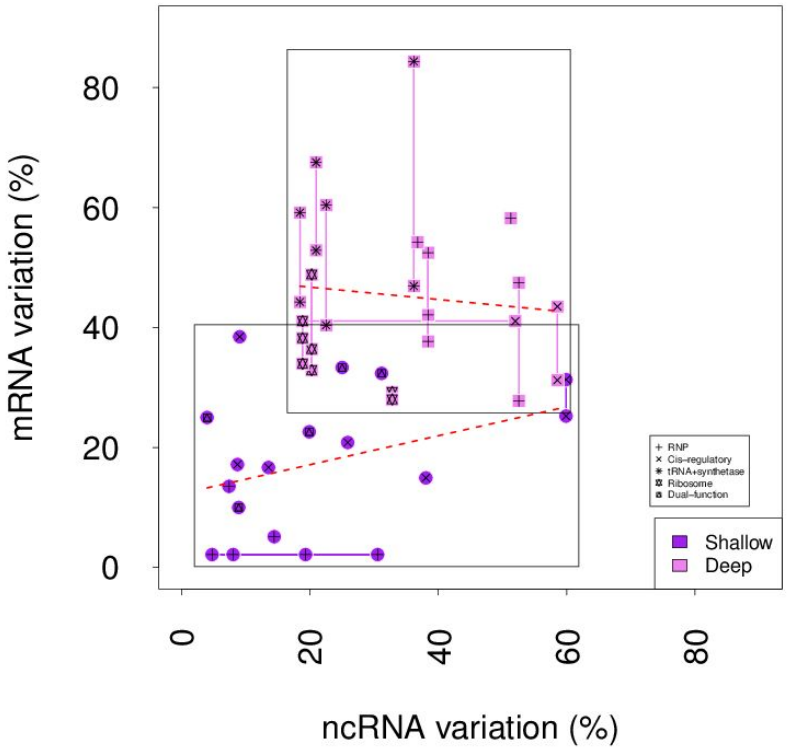

Figure S2: Level of nucleotide variation in interacting RNA and protein pairs.
Interacting RNA and protein pairs were compared across *E. coli* and *S. enterica* (Shallow) and *E. coli* and *N. meningitidis* (Deep). The nucleotide sequence for each RNA-protein pair was scored as the percentage of their respective nucleotide sequence that varied between the species, which gave the percent nucleotide variation. Partners were grouped according to function. No overall correlation (Spearman's correlation) was evident between the variation in one molecule and its interacting partner is evident in either the Deep (*Rho* = -0.08, $p = 0.70$) or Shallow (*Rho* = 0.35, $p = 0.16$) groups, as calculated by Spearman's Correlation.

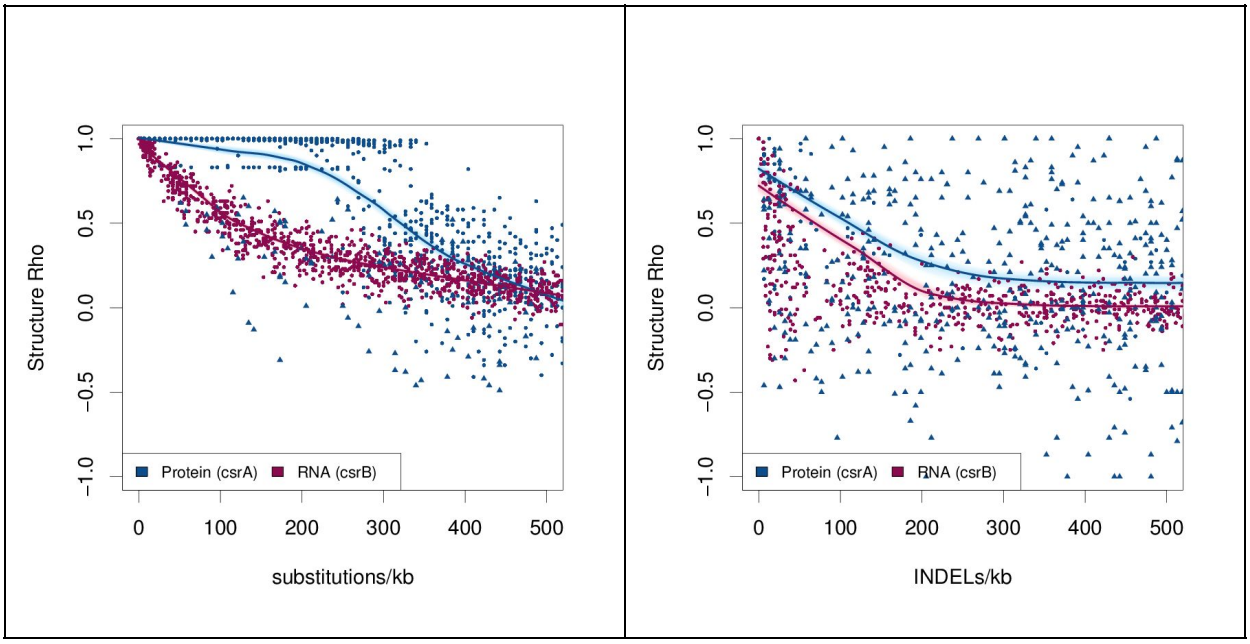

Figure S3: Structural robustness of the CsrA protein and CsrB sRNA. Random mutants of the CsrA messenger RNA (blue) and CsrB small RNA (pink) were generated *in silico*. Their secondary structure probabilities were predicted using "RNAfold-p" and "PSSpred". The per-residue probabilities of either base-paired/not-base-paired or alpha/beta/coil were compared between native and mutated sequences using Spearman's correlation. This gave a "structure rho", where 1 implies the predicted mutant structure is identical to the predicted parental structure, 0 means there is no correlation, and −1 shows a perfect inverse correlation. (A) Substitution mutations and (B) insertion or deletion mutations (indels) were introduced into the RNA (pink) and protein (blue) at rates ranging from 1 to 500 mutations per kilobase (kb). Points corresponding to truncated protein or small RNA with a length less than 75% that of the original are indicated with a solid triangle, otherwise a solid circle is used. Local polynomial regression (loess) curves were fitted to the RNA and protein points. To indicate the confidence for each loess curve, these were bootstrapped 500 times and plotted in light pink or blue to resampled points.

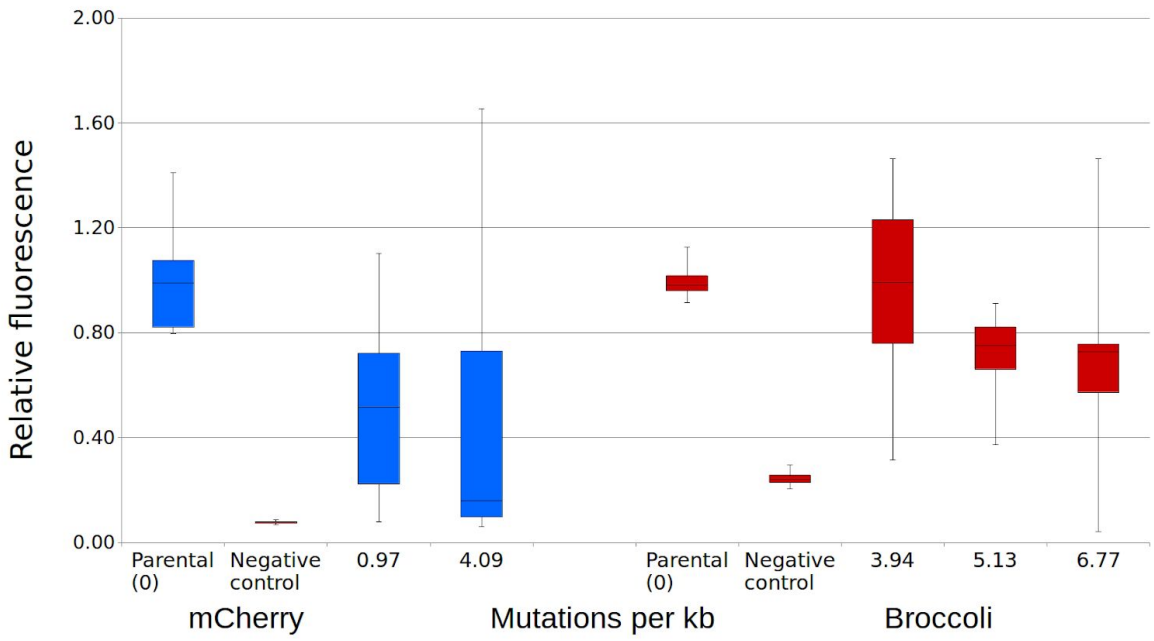

Figure S4: Fluorescence of mutant libraries of RNA aptamer Broccoli and protein mCherry. Libraries of randomly mutated fluorescent RNA aptamer Broccoli and fluorescent protein mCherry were tested for function relative to the unmutated control. Two libraries of mCherry and three libraries of Broccoli were constructed, with using a range of mutation rates per kilobase (kb). The fluorescence intensities of the mutants were normalized to the optical density and the fluorescence intensities of unmutated controls.

References

1. Wassarman KM (2002) Small RNAs in bacteria: diverse regulators of gene expression in response to environmental changes. *Cell* 109(2):141–144.

2. Bartel DP, Chen CZ (2004) Micromanagers of gene expression: the potentially widespread influence of metazoan microRNAs. *Nat Rev Genet*. Available at: http://www.nature.com/nrg/journal/v5/n5/abs/nrg1328.html.

3. Jore MM, et al. (2011) Structural basis for CRISPR RNA-guided DNA recognition by Cascade. *Nat Struct Mol Biol* 18(5):529–536.

4. Fischer S, et al. (2012) An archaeal immune system can detect multiple protospacer adjacent motifs (PAMs) to target invader DNA. *J Biol Chem* 287(40):33351–33363.

5. Marraffini LA (2015) CRISPR-Cas immunity in prokaryotes. *Nature* 526(7571):55–61.

6. Smith AM, Fuchs RT, Grundy FJ, Henkin TM (2010) Riboswitch RNAs: regulation of gene expression by direct monitoring of a physiological signal. *RNA Biol* 7(1):104–110.

7. Breaker RR (2010) Riboswitches and the RNA World. *Cold Spring Harb Perspect Biol* 4(2):a003566–a003566.

8. Gilbert W (1986) Origin of life: The RNA world. *Nature* 319(6055).

9. Hoeppner MP, Gardner PP, Poole AM (2012) Comparative analysis of RNA families reveals distinct repertoires for each domain of life. *PLoS Comput Biol* 8(11):e1002752.

10. Lindgreen S, et al. (2014) Robust identification of noncoding RNA from transcriptomes requires phylogenetically-informed sampling. *PLoS Comput Biol* 10(10):e1003907.

11. de Visser JAGM, et al. (2003) Perspective: Evolution and detection of genetic robustness. *Evolution* 57(9):1959–1972.

12. van Nimwegen E, Crutchfield JP, Huynen M (1999) Neutral evolution of mutational robustness. *Proceedings of the National Academy of Sciences* 96(17):9716–9720.

13. Fontana W, et al. (1993) RNA folding and combinatory landscapes. *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics* 47(3):2083–2099.

14. Kimura M (1984) *The Neutral Theory of Molecular Evolution* (Cambridge University Press).

15. Leontis NB, Stombaugh J, Westhof E (2002) The non‐Watson–Crick base pairs and their associated isostericity matrices. *Nucleic Acids Res* 30(16):3497–3531.

16. Varani G, McClain WH (2000) The G·U wobble base pair. *EMBO Rep* 1(1):18–23.

17. Alkatib S, et al. (2012) The contributions of wobbling and superwobbling to the reading of the genetic code. *PLoS Genet* 8(11):e1003076.

18. Wagner A (2013) *Robustness and Evolvability in Living Systems*.

19. Goldberg AL, Wittes RE (1966) Genetic Code: Aspects of Organization. *Science* 153(3734):420–424.

20. Dayhoff MO, Schwartz RM, Orcutt BC (1978) 22 a model of evolutionary change in proteins. *Atlas of protein sequence and structure*:345–352.

21. Alff-Steinberger C (1969) The genetic code and error transmission. *Proc Natl Acad Sci U S A* 64(2):584–591.

22. Haig D, Hurst LD (1991) A quantitative measure of error minimization in the genetic code. *J Mol Evol* 33(5):412–417.

23. Geyer R, Madany Mamlouk A (2018) On the efficiency of the genetic code after frameshift mutations. *PeerJ* 6:e4825.

24. Maquat LE (1996) Defects in RNA splicing and the consequence of shortened translational reading frames. *Am J Hum Genet* 59(2):279–286.

25. Wang G-S, Cooper TA (2007) Splicing in disease: disruption of the splicing code and the decoding machinery. *Nat Rev Genet* 8(10):749–761.

26. Hindorff LA, et al. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences* 106(23):9362–9367.

27. MacArthur DG, et al. (2014) Guidelines for investigating causality of sequence variants in human disease. *Nature* 508(7497):469–476.

28. Chen J-Q, et al. (2009) Variation in the ratio of nucleotide substitution and indel rates across genomes in mammals and bacteria. *Mol Biol Evol* 26(7):1523–1531.

29. Ohta T (1973) Slightly deleterious mutant substitutions in evolution. *Nature* 246(5428):96–98.

30. Chiu DK, Kolodziejczak T (1991) Inferring consensus structure from nucleic acid sequences. *Comput Appl Biosci* 7(3):347–352.

31. Stombaugh J, Zirbel CL, Westhof E, Leontis NB (2009) Frequency and isostericity of RNA base pairs. *Nucleic Acids Res* 37(7):2294–2312.

32. Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* 89(22):10915–10919.

33. Guo HH, Choe J, Loeb LA (2004) Protein tolerance to random amino acid change. *Proc Natl Acad Sci U S A* 101(25):9205–9210.

34. Kun A, Santos M, Szathmáry E (2005) Real ribozymes suggest a relaxed error threshold. *Nat Genet* 37(9):1008–1011.

35. Mimouni NK, Lyngsø RB, Griffiths-Jones S, Hein J (2009) An analysis of structural influences on selection in RNA genes. *Mol Biol Evol* 26(1):209–216.

36. Abrusán G, Marsh JA (2016) Alpha Helices Are More Robust to Mutations than Beta Strands. *PLoS Comput Biol* 12(12):e1005242.

37. Sanjuan R (2006) In Silico Predicted Robustness of Viroids RNA Secondary Structures. I. The Effect of Single Mutations. *Mol Biol Evol* 23(7):1427–1436.

38. Ferrada E, Wagner A (2012) A Comparison of Genotype-Phenotype Maps for RNA and Proteins. *Biophys J* 102(8):1916–1925.

39. Greenbury SF, Schaper S, Ahnert SE, Louis AA (2016) Genetic Correlations Greatly Increase Mutational Robustness and Can Both Reduce and Enhance Evolvability. *PLoS Comput Biol* 12(3):e1004773.

40. Babajide A, Hofacker IL, Sippl MJ, Stadler PF (1997) Neutral networks in protein space: a computational study based on knowledge-based potentials of mean force. *Fold Des* 2(5):261–269.

41. Ahnert SE (2017) Structural properties of genotype-phenotype maps. *J R Soc Interface* 14(132). doi:10.1098/rsif.2017.0275.

42. You M, Jaffrey SR (2015) Structure and Mechanism of RNA Mimics of Green Fluorescent Protein. *Annu Rev Biophys* 44:187–206.

43. Filonov GS, Moon JD, Svensen N, Jaffrey SR (2014) Broccoli: rapid selection of an RNA mimic of green fluorescent protein by fluorescence-based selection and directed evolution. *J Am Chem Soc* 136(46):16299–16308.

44. Shimomura O, Johnson FH, Saiga Y (1962) Extraction, purification and properties of aequorin, a bioluminescent protein from the luminous hydromedusan, Aequorea. *J Cell Comp Physiol* 59:223–239.

45. Prendergast FG, Mann KG (1978) Chemical and physical properties of aequorin and the green fluorescent protein isolated from Aequorea forskalea. *Biochemistry* 17(17):3448–3453.

46. Tsien RY (1998) The green fluorescent protein. *Annu Rev Biochem* 67:509–544.

47. Felsenstein J (1984) DNAML in PHYLIP 2.6. University of Washington, Seattle.

48. Eddy SR, Bateman A, Finn RD, Petrov AI (2017) Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic acids*. Available at: https://academic.oup.com/nar/article-abstract/46/D1/D335/4588106.

49. Finn RD, et al. (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res* 44(D1):D279–85.

50. Ochman H, Elwyn S, Moran NA (1999) Calibrating bacterial evolution. *Proc Natl Acad Sci U S A* 96(22):12638–12643.

51. Long M, Betrán E, Thornton K, Wang W (2003) The origin of new genes: glimpses from the young and old. *Nat Rev Genet* 4(11):865–875.

52. Cooper VS, Vohr SH, Wrocklage SC, Hatcher PJ (2010) Why genes evolve faster on secondary chromosomes in bacteria. *PLoS Comput Biol* 6(4):e1000732.

53. Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW (2002) Evolutionary rate in the protein interaction network. *Science* 296(5568):750–752.

54. Wadler CS, Vanderpool CK (2007) A dual function for a bacterial small RNA: SgrS performs base pairing-dependent regulation and encodes a functional polypeptide. *Proc Natl Acad Sci U S A* 104(51):20454–20459.

55. Bobrovskyy M, Vanderpool CK (2014) The small RNA SgrS: roles in metabolism and pathogenesis of enteric bacteria. *Front Cell Infect Microbiol* 4:61.

56. Zhang Y PSSpred: A multiple neural network training program for protein secondary strucure prediction. Available at: http://zhanglab.ccmb.med.umich.edu/PSSpred [Accessed July 20, 2018].

57. Zuker M, Stiegler P (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res* 9(1):133–148.

58. McCaskill JS (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* 29(6-7):1105–1119.

59. Rogers E, Murrugarra D, Heitsch C (2017) Conditioning and Robustness of RNA Boltzmann Sampling under Thermodynamic Parameter Perturbations. *Biophys J* 113(2):321–329.

60. Mattick JS, Makunin IV (2006) Non-coding RNA. *Hum Mol Genet* 15 Spec No 1:R17–29.

61. Hershberg R, Altuvia S, Margalit H (2003) A survey of small RNA-encoding genes in Escherichia coli. *Nucleic Acids Res* 31(7):1813–1820.

62. Drummond OE (2005) Tracking and classification with attribute data from multiple legacy sensors. *Signal and Data Processing of Small Targets 2005* doi:10.1117/12.624910.

63. Nawrocki EP, Eddy SR (2007) Query-dependent banding (QDB) for faster RNA similarity searches. *PLoS Comput Biol* 3(3):e56.

64. Eddy SR, Durbin R (1994) RNA sequence analysis using covariance models. *Nucleic Acids Res* 22(11):2079–2088.

65. Carothers JM, Oestreich SC, Davis JH, Szostak JW (2004) Informational complexity and functional activity of RNA structures. *J Am Chem Soc* 126(16):5130–5137.

66. Omer S, Harlow TJ, Gogarten JP (2017) Does Sequence Conservation Provide Evidence for Biological Function? *Trends Microbiol* 25(1):11–18.

67. Bloom JD, Labthavikul ST, Otey CR, Arnold FH (2006) Protein stability promotes evolvability. *Proc Natl Acad Sci U S A* 103(15):5869–5874.

68. Wuchty S, Fontana W, Hofacker IL, Schuster P (1999) Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers* 49(2):145–165.

69. Li C, Qian W, Maclean CJ, Zhang J (2016) The fitness landscape of a tRNA gene. *Science* 352(6287):837–840.

70. Rost B (1999) Twilight zone of protein sequence alignments. *Protein Eng* 12(2):85–94.

71. Gardner PP, Wilm A, Washietl S (2005) A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucleic Acids Res* 33(8):2433–2439.

72. Goldstein RA (2008) The structure of protein evolution and the evolution of protein structure. *Curr Opin Struct Biol* 18(2):170–177.

73. Paige JS, Wu KY, Jaffrey SR (2011) RNA mimics of green fluorescent protein. *Science* 333(6042):642–646.

74. Autour A, Westhof E, Ryckelynck M (2016) iSpinach: a fluorogenic RNA aptamer optimized for in vitro applications. *Nucleic Acids Res* 44(6):2491–2500.

75. Dolgosheina EV, et al. (2014) RNA mango aptamer-fluorophore: a bright, high-affinity complex for RNA labeling and tracking. *ACS Chem Biol* 9(10):2412–2420.

76. Introduction to Fluorescent Proteins *Nikon's MicroscopyU*. Available at: https://www.microscopyu.com/techniques/fluorescence/introduction-to-fluorescent-proteins [Accessed August 10, 2018].

77. Ellington AD, Szostak JW (1990) In vitro selection of RNA molecules that bind specific ligands. *Nature* 346(6287):818–822.

78. Arnold FH (1998) Design by Directed Evolution. *Acc Chem Res* 31(3):125–131.

79. Jakočiūnas T, Pedersen LE, Lis AV, Jensen MK, Keasling JD (2018) CasPER, a method for directed evolution in genomic contexts using mutagenesis and CRISPR/Cas9. *Metab Eng* 48:288–296.

80. Fontana W, Schuster P (1998) Continuity in evolution: on the nature of transitions. *Science* 280(5368):1451–1455.

81. Schuster P (2001) Evolution in silico and in vitro: the RNA model. *Biol Chem* 382(9):1301–1314.

82. McBride RC, Ogbunugafor CB, Turner PE (2008) Robustness promotes evolvability of thermotolerance in an RNA virus. *BMC Evol Biol* 8:231.

83. Lenski RE, Barrick JE, Ofria C (2006) Balancing robustness and evolvability. *PLoS Biol* 4(12):e428.

84. Nawrocki EP, Eddy SR (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* 29(22):2933–2935.

85. Eddy SR (2011) Accelerated Profile HMM Searches. *PLoS Comput Biol* 7(10):e1002195.

86. Suyama M, Torrents D, Bork P (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res* 34(Web Server issue):W609–12.

87. Yan R, Xu D, Yang J, Walker S, Zhang Y (2013) A comparative assessment and analysis of 20 representative sequence alignment methods for protein structure prediction. *Sci Rep* 3:2619.

88. Yang J, et al. (2015) The I-TASSER Suite: protein structure and function prediction. *Nat Methods* 12(1):7–8.

89. Lorenz R, et al. (2011) ViennaRNA Package 2.0. *Algorithms Mol Biol* 6:26.

90. Miyazaki K (2011) MEGAWHOP cloning: a method of creating random mutagenesis libraries via megaprimer PCR of whole plasmids. *Methods Enzymol* 498:399–406.