The genetic robustness of RNA and protein from evolutionary, structural and functional perspectives.

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. Please address correspondence to Dorien Coray (doriencoray@gmail.com) and Paul Gardner (paul.gardner@otago.ac.nz).

Abbreviations:
BANP: Basic, Acidic, Non-polar, Polar; CRISPR: Clustered Regularly Interspaced Palindromic Repeats; INDELs: Insertions and Deletions; ncRNA: Non-coding RNA; ORF: Open Reading Frame; RNase P: Ribonuclease P; RNP: ribonucleoprotein; rRNA: ribosomal RNA; tRNA: transfer RNA; RY: purine-pyrimidine

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 despite mutations. 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 test the relative mutational robustness of RNA and protein pairs using three approaches: evolutionary, structural, and functional. We compare the nucleotide diversities of functional RNAs with those of matched proteins. Across different levels of conservation, we found the nucleotide-level variations between the biomolecules largely overlapped, with proteins generally supporting more variation than matched RNAs. 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, yet again proteins are slightly more robust on average than RNAs. *In vitro*, mutated fluorescent RNAs retained greater levels of function than the proteins. Overall this suggests that while protein has slightly higher robustness than RNA as a group, neither RNA and protein is consistently more robust than the other. 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 non-coding 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 influencing 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

RNA & Protein: RNA is mainly known for its role in translation, yet it is also involved in controlling gene expression (e.g., bacterial small RNAs and microRNAs), intrinsic immunity (e.g., CRISPR-mediated acquired immunity) and the cell's response to environmental stimuli (e.g., thermosensors and riboswitches) [1,2]. The discovery of functional RNAs, like transfer RNA and catalytic RNAs, led to the proposal of an ancestral RNA world, where RNA both catalyzes the reactions of life and encodes genetic information [3]. Despite the importance of non-coding RNAs (ncRNAs) to cellular function, many RNAs exhibit low sequence conservation and are not as broadly distributed as their proteinaceous brethren [4]. In contrast, proteins make up the bulk of the biomolecular contents of the cell, and are required for the majority of critical cellular structures and functions. Yet protein production is a complex, multi-stage process that is sensitive to errors [5–7].

Robustness: The ability to preserve a phenotype in the face of sequence perturbations is termed mutational robustness [8–11]. More robust molecules maintain their phenotypes despite mutations, while less robust molecules lose their function rapidly. The genotype level at which mutations are considered, and the types of phenotypes (structure, function) that are measured, can vary between analyses. Here, we consider mutations in nucleotide sequence and how they modify phenotypes such as the structure and function of encoded genes. The robustness of individual RNA and protein molecules are both affected by factors such as the stability of individual molecules, interactions and gene expression levels [12,13]. For example, RNA stems are generally more robust than loops [14–16] and allow non-canonical base-pairs [17] while protein alpha helices are more robust than beta strands and both were more robust than unstructured coils [18,19]. RNAs and proteins also have independent mechanisms that allow for near-neutral changes that maintain molecular structures and functions (Table 1).

Primary Structure: 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 with interchangeable tRNAs [20,21]. 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 [22]. Monte Carlo simulations have shown that the extant genetic code is significantly more robust to substitution and frameshift mutations than randomly generated genetic codes [23–28]. Premature stop codons and frameshift errors can be introduced by substitutions or indels (insertions and deletions). While ncRNA production requires transcription—and in some cases, additional maturation such as editing, splicing or processing—proteins require these in addition to translation, which depends upon the maintenance of a correct reading frame [29]. These additional steps likely amplify the potential harm of nucleotide changes [5]. This is supported by the fact that disease-associated sequence variation is enriched ten-fold in human protein-coding regions [6,30] and that overall variation is reduced in coding regions, particularly indels [7]. 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 [26,31]

Mechanism	RNA	Protein
Primary sequence	Preponderance of transition mutations (R to R, Y to Y) over transversions (R to Y) [26]	Maintenance of amino acid sequence through degenerate genetic code e.g. CGN = Arg, GGN = Gly [20]
Secondary/ tertiary structure	Non-canonical base-pairing maintains stems and loops, e.g., wobble base-pairing (G:U), covarying sites may preserve base-pairs [32], isostericity of non-canonical base-pair interactions [17].	Point mutations may preserve the biochemistry of amino acids, maintaining structure/function, e.g., hydrophobic residues stay hydrophobic [22,33]
Size	Shorter molecules are more robust as they are less likely to acquire mutations at a fixed rate of mutations per kb [19,34].	
Stability	More stable structures may buffer mutations, e.g., Stems > loops for RNA [14–16] and alpha helices >beta sheets>loops for protein [18,19].	

Previous robustness studies: Previous comparisons of protein and RNA 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) [26,35–38], found that RNA networks differed from protein networks, being larger (more robust) but also less compact. More recent work has shown that this is dependent on the mathematical framework used, and suggested that RNAs and proteins are more similar [37,38].

To compare protein and RNA robustness on as fairly as possible we have designed a series of experiments that explore the question of which is more robust. 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 any truly fair comparisons. Therefore we have carefully selected

matched pairs of proteins and RNAs that share a function or structure, and compare their evolutionary conservation over matched time periods, their structural robustness with simulated mutations and their functional robustness with error-prone PCR. In exploring the broad differences between protein and RNA robustness under comparable conditions, we may gain insight into the origins and bioengineering potential of RNAs and proteins.

Results

We have devised three tests, two *in silico* and one *in vitro,* to explore whether RNAs and proteins differ in their robustness to mutation. **1.** We have considered the degree of DNA sequence variation between groups of RNA and protein families over matched time-periods. To control variation as much as possible, ncRNA genes were compared to the genes of proteins that are involved in related functions (e.g., ribosomal RNAs and proteins, riboswitches and the proteins these regulate, tRNAs and tRNA synthetases, etc.). These RNA-protein pairs have shared recent phylogenetic 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. See Figure 1. **2.** We have simulated point mutations and insertions/deletions in DNA sequences encoding functionally linked proteins and ncRNAs, these are chiefly ribonucleoparticles (RNPs) with solved tertiary structures. The potential impacts of mutations have then been assessed based upon structural stability ($\Delta\Delta G$), evolutionary profiles (Δ bitscore) and predicted secondary structure profiles. See Figure 2. **3.** We have compared the functional robustness of a protein and RNA directly with mutagenesis of a functionally comparable fluorescent RNA [39,40] and fluorescent protein [41–43], and quantified the fraction of mutants that maintained function for each class of molecule. See Figure 3.

Experiment 1: Sequence diversities of ncRNAs and proteins involved in the same processes

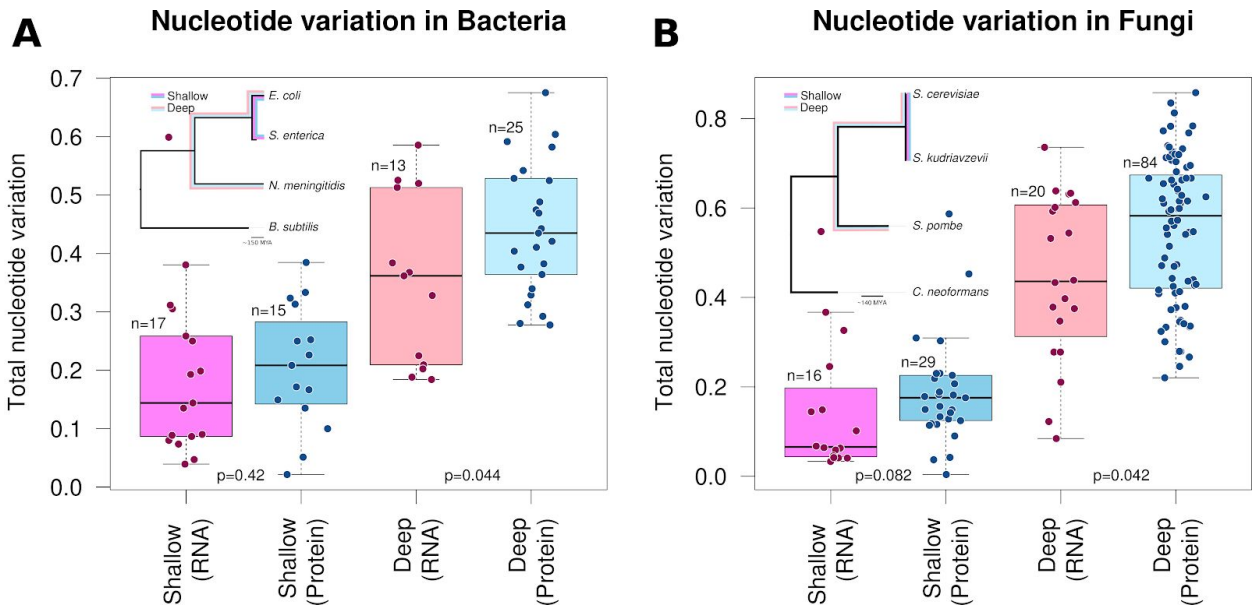


Figure 1: Proportion of variable nucleotides between shallow and deep divergence times for conserved RNA and protein families. The proportion of total nucleotide variation was calculated for aligned orthologous RNAs (pink) and proteins (blue) and illustrated with box-whisker and jitter-plots. (**A**) RNAs and proteins from representative bacteria, *Escherichia coli* and *Neisseria meningitidis* (deep divergence, lighter shades) or *E. coli* and *Salmonella enterica* (shallow divergence, darker shades). **(B)** RNAs and proteins from representative fungi, *Saccharomyces cerevisiae* and *Schizosaccharymocyes pombe* (deep divergence, lighter shades) or *S. cerevisiae* and *Saccharomyces kudriavzevii* (shallow divergence, darker shades). **Tree inserts:** The relationship between the deep (lighter shades) and shallow (darker shades) diverged species on SSU rRNA, (`dnaml`) [44], phylogenetic trees. It should be noted that the 'shallow' species are estimated to have diverged less than 150 million years ago (MYA) in each case, whereas the 'deep' species diverged approximately >400 million MYA in each case.

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, and are found in the Figure 1 reference genomes [45,46]. These can be broadly classified as ribonucleotide particles (RNPs), cis-regulatory elements and corresponding proteins, and dual-function genes where one partner is modified or processed by the other (Tables S1 & S2). 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 **four** independent sets of RNA:protein pairs in yeast and bacteria species that were divergent enough to exhibit sequence diversity and have matched G+C contents (~50%) (Figure 1). Two sets are for "deep" gene pairs that are highly conserved, often involved in functions such as transcription or translation. The species compared have diverged more than 400 million years ago (*Saccharomyces cerevisiae* and *Schizosaccharymocyes pombe* for yeast, *Escherichia coli* and *Neisseria meningitidis* for

bacteria). The other two sets are for "shallow" gene pairs that have diverged recently. These gene-pairs generally have a limited phylogenetic distributions, and possibly a high tolerance of mutations that is a characteristic of new genes [47,48], which are often involved in functions such as gene regulation, biosynthesis and maturation of RNPs. The species compared have diverged less than 150 million years ago (*S. cerevisiae* and *Saccharomyces kudriavzevii* for yeast, *E. coli* and *Salmonella enterica* for bacteria).

At each level of conservation, we computed the total number of variable nucleotide sites between the two species for a given RNA and protein domain. This was normalized by the length of the sequence to derive the total nucleotide variation for each. Variants were further classified as neutral or not, depending on whether secondary structure (RNA), amino acids (protein) and/or the biochemistry of the RNA or protein was preserved (Figure S1).

Generally, the level of total nucleotide variation within protein-coding sequences was higher than the RNAs, yet the distributions for both largely overlapped. Only the distributions for just the deep gene pairs could be considered to be statistically significantly different ($p=0.044$ in bacteria, and $p=0.042$ in fungi, two-sided Wilcoxon rank-sum test) (Figure 1A&B). While the shallow gene pairs were not statistically significantly different from each other ($p=0.42$ in bacteria, and $p=0.082$ in fungi, two-sided Wilcoxon rank-sum test) (Figure 1A&B).

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 [49]. We did not, however, see a general significant 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. While protein-coding mRNAs tend to have greater nucleotide diversity than RNAs it is typically not significantly higher over short time-scales, yet is marginally significant over longer timescales. By this measure, RNAs and proteins exhibit comparable degrees of robustness.

Experiment 2: simulated mutations and the predicted impacts on structural and evolutionary robustness

We have simulated point mutations and insertions/deletions in a number of selected pairs of protein and RNA encoding DNA sequences. The impacts of simulated mutations have been analysed using tools for assessing mutation impacts on structure or function that work on both protein and RNA.

For the first simulation we use non-redundant RNPs with solved structures (e.g. tRNA-synthetase pairs, signal recognition particles, 6S RNA bound by RNA polymerase, components of the spliceosome, a guide RNA in a complex with a Cas protein, and snoRNAs with associated proteins) available in the PDB (July 2019) [50]. We have inferred the theoretical impact of point mutations on structures by estimating

changes in Gibbs free energy, ($\Delta\Delta G$, units: kcals/mol) [51,52] (Figure 2B). The variants have been further analysed using an orthogonal approach for assessing evolutionary robustness using differences in profile HMM scores (Δ bitscore, units: bits) [53,54] (Figure 2A). Each mutant sequence could be classed as divergent from the parental (zero) if the corresponding Δ bitscore value is less than zero, or as destabilised if the corresponding $\Delta\Delta G$ is greater than zero. We found that the proportion of ncRNA mutants that were classed as either divergent or destabilised was consistently greater than that for proteins (see Table S1 for the numerical values).

In addition to the above, we have conducted a more detailed analysis of secondary structure robustness on a dual-function protein/ncRNA pair SgrS/SgrT, where the correlation between predicted probabilities of secondary structure elements have been compared between native and mutated sequences [55]. SgrS and SgrT were selected as they are both structured and short (for computational efficiency) and the protein and RNA structures are not contained in the Protein Data Bank (PDB) that the secondary structure prediction tools we use were trained on (e.g. PSSpred).

We compute per-residue secondary structure probabilities on native simulated point or insertion/deletion mutant sequences. The correlation between native and mutant per-residue secondary structure probabilities is computed ("Structure Rho" i.e. Spearman's correlation coefficient) and the distributions of these for a range of mutation densities have been plotted (Figure 2C&D). Both the protein and RNA mutants retain about half the parental structure (structure correlation of 0.5) at 100 point mutations per kilobase, though the protein shows a higher variance in response to mutation compared to RNA (Figure 2C&D). The protein is slightly more sensitive to indels than RNA, but shows a very similar overall level of decline in its structure.

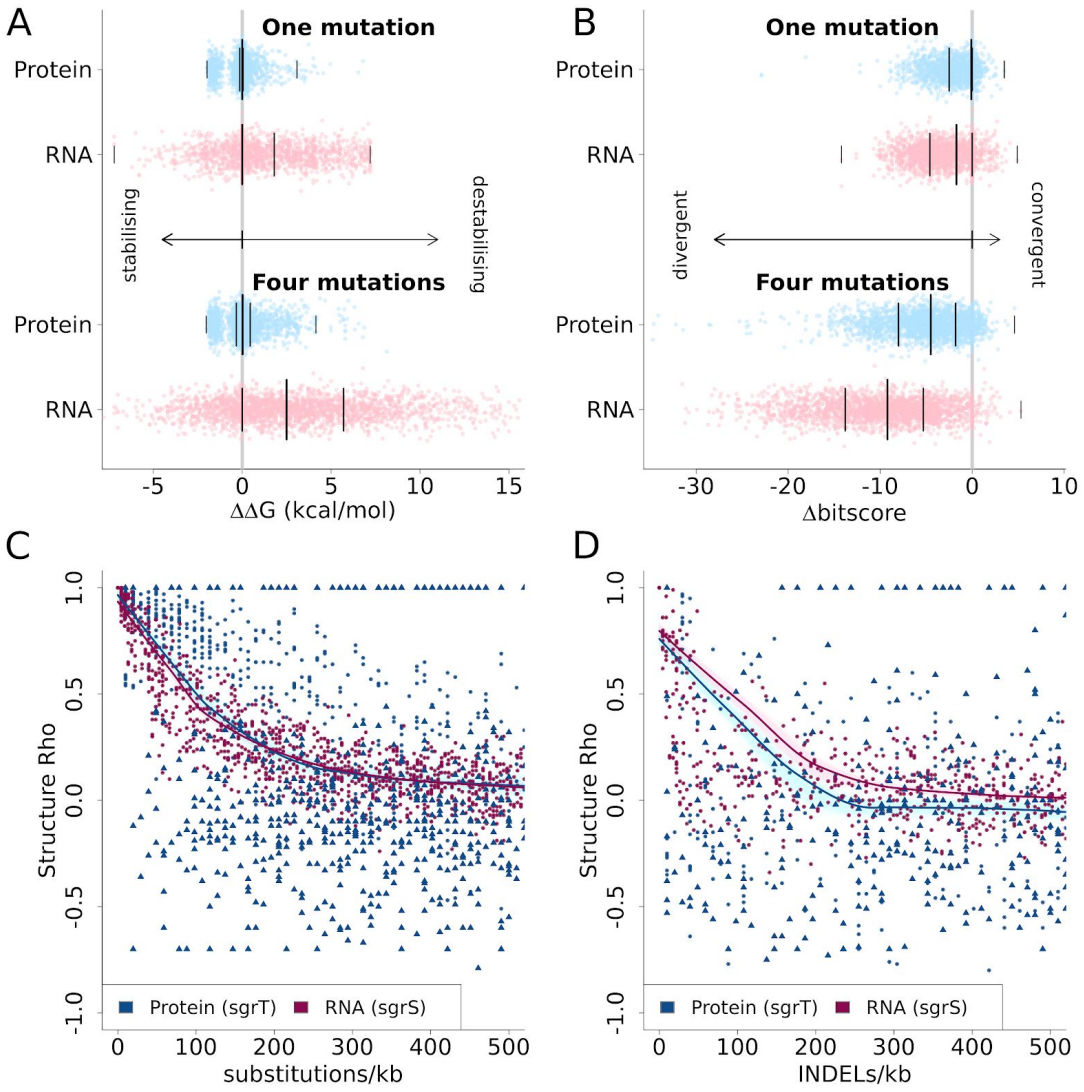


Figure 2: Robustness of structure predictions to random *in silico* mutagenesis for selected protein-RNA pairs. (**A&B**) For 24 non-redundant pairs of RNP structures available in the PDB, either one (upper) or four (lower) point mutations have been randomly introduced into corresponding DNA sequences. The impact of these simulated mutations on protein and RNA sequences have been assessed using predicted Δ bitscore (**A**) and $\Delta\Delta G$ (**B**) values. The distributions of values are illustrated using jitter-plots, with the medians, 25th and 75th percentiles, and extremes indicated with a vertical line for each distribution. (**C&D**) Comparing the robustness of protein and RNA secondary structures to random substitutions or insertion/deletion (INDEL) mutations. Secondary structure probabilities were predicted for native and mutated sequences, and 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 (Structure Rho). Truncated proteins or sRNAs with a length less than 75% of the original are indicated with a solid triangle, otherwise a solid circle is used. Truncated regions and other unalignable residues were excluded from the correlation calculation. 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 (RNA) or blue (protein) to resampled points.

Experiment 3: Mutational robustness of a functionally equivalent RNA and protein

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). To test for robustness of function, we have mutated an RNA and a protein matched for an assayable function (fluorescence) and tested these mutations *in vivo*.

To investigate how biomolecules may differ in their robustness to mutations in DNA, we constructed mutant libraries of the fluorescent RNA aptamer Broccoli [39,40] and the fluorescent protein mCherry [43]. 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 are similar, with few insertions or deletions (indels) and similar numbers of transitions and transversions.

A

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

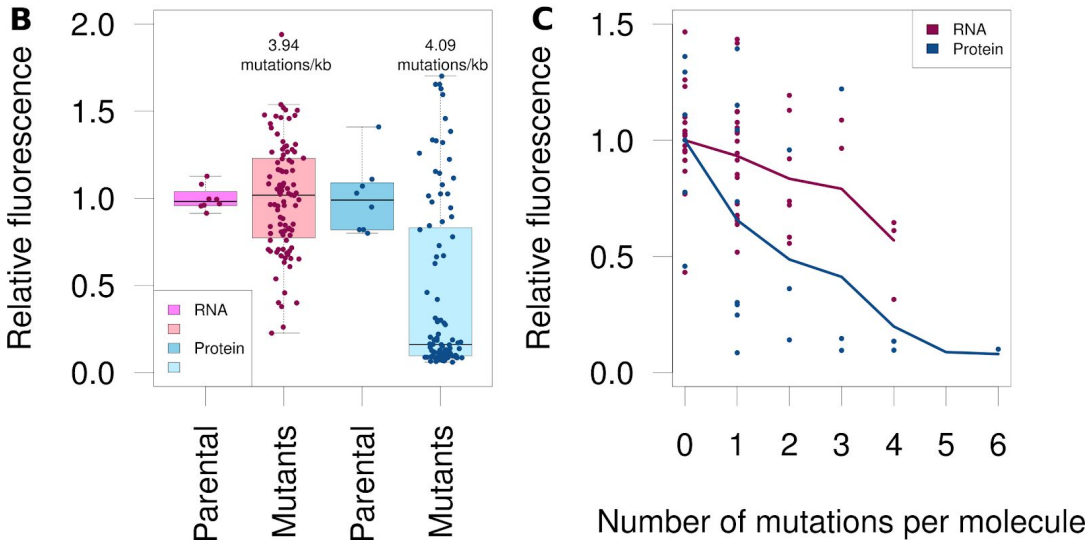


Figure 3: Relative fluorescence intensities of mutated RNA Broccoli and mutated protein mCherry. Libraries of randomly mutated fluorescent RNA aptamer Broccoli and fluorescent protein mCherry have been generated and tested for function relative to an unmutated control. (A) The mCherry and Broccoli libraries 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.

Discussion

Summary: We have hypothesized that RNA would be more robust than proteins. This is supported by the fact that RNA often requires less processing than protein to produce functional molecules [56], is not susceptible to frameshift mutations [57] and is less likely to be found over broad evolutionary distances with homology searches (possibly due to high higher mutation rates) [4]. Our multi-scale tests of RNA and protein robustness revealed no consistent evidence to support our hypothesis. This leads us to conclude that both molecules are remarkably robust to mutation, and that proteins are likely to be slightly more robust than RNAs based upon the balance of all the evidence. This finding corroborates some previous work, suggesting that RNAs and proteins have similar overall robustness to mutation [37,38].

Experiment 1: Our investigation of RNA and protein robustness was, in part, initiated by the observation that RNAs and proteins are differentially distributed across phylogenetic distances [4]. A possible explanation for the difference is a higher mutation rate for functional RNAs, making them difficult to detect. However, we did not find any evidence to support this possibility when comparing the total nucleotide variation of matched RNAs and proteins over matched evolutionary divergences. If RNAs are not more robust than proteins, as our experiments imply, other factors must account for the apparent differences in phylogenetic distributions. For example, it is likely that protein homology search is statistically more powerful than that for nucleotides [58,59]. It could also be speculated that gene turnover and neofunctionalization are more rapid for RNAs than for proteins [48].

Experiment 2: 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. Comparing the sequence diversity of extant RNAs and proteins revealed similar patterns in bacteria and fungi. Rates of nucleotide substitution were close, with proteins generally having more nucleotide diversity than RNAs. This was also seen when comparing Δ bitscore and $\Delta\Delta$ G of RNPs mutated *in silico*. The proportion of divergent/destabilised ncRNAs was similar but higher than that seen with proteins. A specific protein-RNA pair, SgrS/SgrT was tested more in depth to both point mutations and indels. The RNA was slightly more robust than proteins in response to indels, however the overall responses were similar in both molecule types. These measures of RNPs are proxies of robustness, but taken together suggest that RNA is no more robust than proteins. They responded remarkably similarly but protein appears to be somewhat more robust than RNA.

The genetic robustness of proteins, especially to indels, is unexpected for a few reasons. The need for translation increases the avenues of error in protein production, and the evolution of translational robustness has been considered a factor in constraining nucleotide diversity in highly expressed proteins [60]. While INDELs can greatly change downstream amino acids through frameshifts, RNA has no code to protect, and can, in theory, absorb additional nucleotides in stem bulges [15,61,62]. Nonetheless, the predicted structures for RNA SgrS were almost as sensitive to indels as the protein SgrT.

Experiment 3: The functional test of the fluorescent RNA Broccoli and protein mCherry was the only test where the RNA was considerably more robust. In order to match the mRNA and ncRNA sequence lengths as closely as possible we used a double Broccoli aptamer, which means a disabling mutation in one half can leave the other half with functionality. Consequently, we may have inadvertently increased the robustness of the RNA over the protein.

Future work: Future evolutionary, theoretical and experimental studies are needed to explore the initial results presented here. We acknowledge the limitations of using just one RNA and protein pair for analysis, the mutagenesis could be repeated with different fluorescent RNAs (e.g., Spinach [63], iSpinach [64], Mango [65]) and phylogenetically distinct fluorescent proteins (e.g., GFP, luciferase, ZsGreen1, ZsYellow1 [66]), for example. Furthermore, simulated evolution experiments such as SELEX for RNAs [67] and directed evolution for proteins [68,69] could be performed to identify differences in RNA and protein evolvability as well as mutational robustness. Alternatively, experimental evolution datasets could be mined for appreciable differences between protein and RNA evolution [70]. Evolution and evolvability may also be modeled computationally using flow reactor simulations [71,72]. Forms of robustness other than mutational robustness, like the robustness of protein and RNA interaction networks, and robustness

to environmental conditions such as temperature [73], pressure and pH fluctuations [74] can also be explored. The study of the interplay between robustness and evolvability informs our understanding of how new functions evolve in proteins and RNAs [37,75–77], the proposed transition from an RNA world [78] and may help engineer biomolecules capable of functioning under mutational and environmental challenges.

Methods

Natural variation of functional ncRNA:protein systems

We have curated pairs of RNA and protein families from Rfam and Pfam that are linked by either direct interactions or by process, and conserved between either *E. coli* and *N. meningitidis* (deep divergence; n=26 pairs) or *E. coli* and *S. enterica* (shallow divergence; n=18 pairs) in bacteria; in fungi we investigated RNPs from, *Saccharomyces cerevisiae* and *Schizosaccharymocyes pombe* (deep divergence; n=98 pairs) or *S. cerevisiae* and *Saccharomyces kudriavzevii* (shallow divergence; n=49 pairs). The RNA and protein coding sequences were aligned using the most accurate techniques appropriate for each molecule type [59,79]. Each pair of deep or shallow diverged nucleotide sequences were aligned, for RNA using Rfam, tRNAscan-SE (v1.3.1) or Intron covariance models [80–82] and `cmalign` (v1.1.1) [83] or, for the protein domains, using `hmmalign` (v3.1b2) [84] and concordant codon-aware nucleotide alignments generated with `PAL2NAL` (v14) [85]. The total number of variant sites was recorded for each alignment, and a nonparametric Mann-Whitney U rank-based test [86] implemented in R was used to compare the distributions of the number of variant sites for protein and RNA. The results can be viewed in Figure 1 and Supplementary Tables 1&2.

The phylogenetic trees embedded in Figures 1A&B are crafted from SSU rRNA alignments. Sequence alignments for the three model bacterial or fungal species and an outgroup were generated by aligning to the corresponding Rfam covariance model [87] (bacterial SSU - RF00177 and eukaryotic SSU - RF01960, for the fungi) with `cmalign` [83], a phylogenetic tree was estimated for each alignment using `dnaml` (v3.69) from the phylip package [44,88] with default parameters.

The corresponding computer code for the above steps is available in '`computeSynonNonsynon.pl`' and '`plotConservation.R`' scripts in the github repository. Raw data is available in the `data/genomes-bacterial/` and the `data/genomes-fungal/` directories.

Simulated variation and ribonucleoprotein secondary structure

A curated list of RNP structures was selected by analysing the RNA sequences available in the PDB sequence repository ftp://ftp.wwpdb.org/pub/pdb/derived_data/pdb_seqres.txt.gz -- (18 July 2019). RNA

sequences were extracted and checked for plausible minimum free energy structures using `RNAfold` from the Vienna RNA Package [89]. RNP selections were based upon the presence of a stable RNA structure and diversity (e.g. avoiding an over-abundance of any particular structure type, or species). In addition, ncRNA lengths between 50 and 150 nucleotides, and protein lengths between 50 and 500 were favoured (although some exceptions have been made).

Messenger RNAs for the selected proteins were recovered by `tblastn` search of the NR database. If necessary, mRNAs were manually corrected to match the corresponding PDB sequences. Quality control steps included verifying that RNA, protein and mRNA sequences were consistent with the corresponding PDB entries. Protein and RNA sequences were further checked for corresponding profile models in the EggNOG database (v5.0) [90] and Rfam (v14.0) [87] using HMMER3 profile HMMs [84] and Infernal covariance models [83] respectively. See Table S3 for a list of the final 24 structures that met our criteria.

In order to compute $\Delta\Delta G$ values (change in ΔG - Gibbs free energy - due to mutations), we selected a machine-learning method, `MAESTRO` [91] for the proteins, and a nearest-neighbour free-energy method, `RNAfold` [89,92] for the RNAs. Missing values due to method failures were recorded as NA's in our output files. `MAESTRO` failed on 15% of the four mutations, and 4% of the one mutation simulations, while `RNAfold` failed on 0.2% of the four mutation simulations. In the worst case, an additional 15% high $\Delta\Delta G$ values for proteins would not have altered our conclusions. In order to compute Δbitscore values (change in profile HMM or CM bitscores due to mutations [53]), we used HMMER3 profile HMMs [84] from the EggNOG database [90] for the proteins and Infernal covariance models [83] from the Rfam database [87] for the RNAs. Missing values due to method failure were again recorded, and in the worst case affected 1.5% of the simulations, again, not enough to alter our main conclusion. The results can be viewed in Figure 2A&B and Supplementary Tables 3&6. The $\Delta\Delta G$ and Δbitscore for proteins and RNAs were only compared qualitatively, since the methods and models to compute each are quite different.

The corresponding computer code for the above steps is available in `computeDeltaDelta.pl` and `plotDelta.R` scripts in the github repository. Raw data is available in the `data/delta-delta/` directory.

The RNA sequences of SgrS RNA (Rfam accession: RF00534, 227 nucleotides long) and corresponding protein SgrT (Pfam accession: PF15894, 102 nucleotides long) were mutated *in silico* with random substitution or indel mutations 100 times, with mutation rates varying between 0 and 500 per kilobase. We used `PSSpred` to infer the probability of each residue in the SgrT mutants forming an alpha helix, beta sheet or coil [93,94]. For the RNA sequences, we used "`RNAfold -p`", an implementation of McCaskill's RNA partition folding function [95] found in the Vienna RNA Package [89]. The results can be viewed in Figure 2C&D.

The corresponding computer code for the above steps is available in `structureMutagenerator.pl` and `plotStructureMutagenerator.R` scripts in the github repository. Raw data is available in the `data/sgr-structural/` directory.

Fluorescent protein and RNA construction and measurements

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) (Figure SX). 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 [96]. 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.

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. A summary of the results can be viewed in Figure 3A-C.

The corresponding computer code for the above steps is available in `plotFluoro.R` script in the github repository. Raw data is available in the `data/fluoro/` directory.

Data and software availability

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

Acknowledgements

We thank the reviewer who challenged the senior author's unsupported statement "RNAs, unlike proteins, are relatively robust to genetic variation" in a 2014 draft manuscript, thus inspiring this study. We thank Gabrielle David, PhD, for editing a draft of this manuscript. We are grateful to Dr Elena Rivas and Professor Dan Tawfik who gave critical feedback on the draft manuscript, and to the attendees of the 2018 Benasque "Computational Approaches to RNA Structure and Function" conference for many valuable discussions. This work was supported by Biomolecular Interaction Centre postdoctoral grants (DS Coray) and by a Rutherford Discovery Fellowship administered by the Royal Society of New Zealand awarded to P. Gardner.

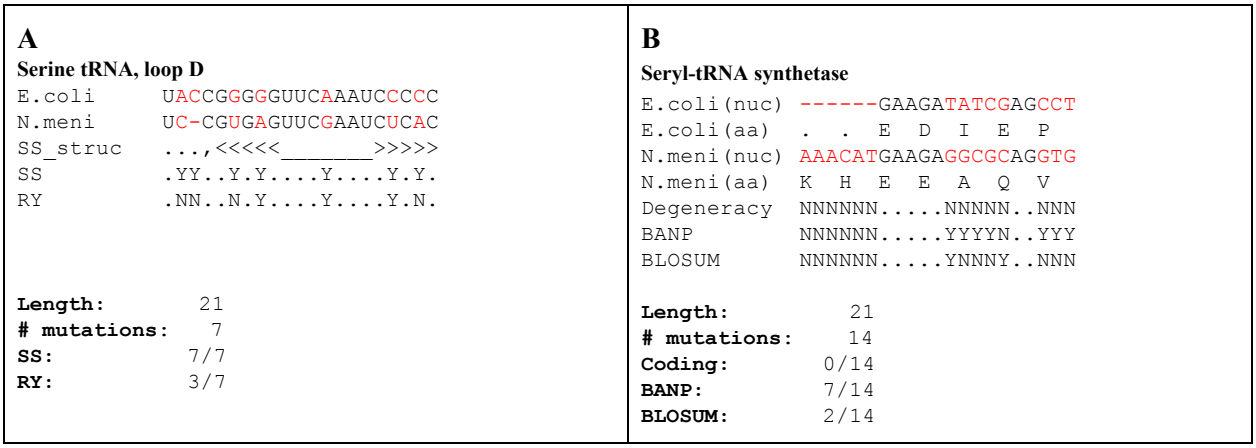
Supplementary Material

Supplementary Tables:

Table S1: Bacterial RNPs and data for Figure 1A
Table S2: Fungal RNPs and data for Figure 1B
Table S3: RNPs for structural analysis for Figure 2A&B
Table S4: Fluorescence data for Figure 3
Table S5: Fluorescent RNA and protein sequences for Figure 3
https://docs.google.com/spreadsheets/d/1exZaYpTQRfTpdNBVaIOJID3Uzw_WIJy0XBOTmPaOSi4/edit?usp=sharing

Comparison of neutral variation between extant sequences in RNPs

Variant sites identified between deeply or recently conserved RNPs (Table S1&S2) were identified and classified as neutral if it preserved structure and/or biochemistry (Figure S1A&B). In RNA secondary structure was preserved more than biochemistry (Figure S1C). In protein on average a third to two-thirds (~0.3 bacteria, ~0.6 fungi) of the variations were synonymous, with many of the non-synonymous variations preserving biochemistry of the coded amino acid according to two measures (Figure S2C). Bacterial protein and RNA had higher proportions of non-neutral variation than fungi.



C

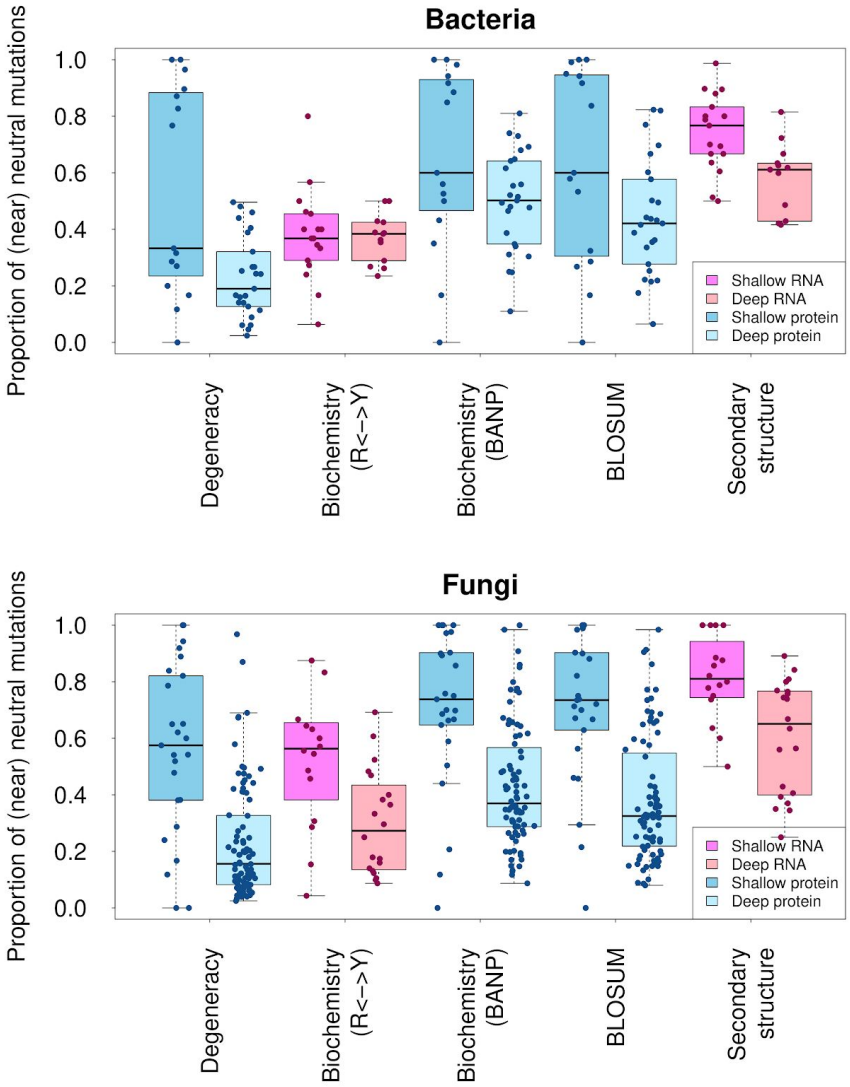


Figure S1: The proportion of nucleotide variants in RNA or protein that can be classed as neutral. A collection of functionally linked RNA and protein families that are shared between *E. coli* and *N. meningitidis* (*N. meni*) (Deep, lighter shades) or between *E. coli* and *S. enterica* (Shallow, darker 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 acid replacements were assigned a non-negative score in the **BLOSUM** score matrix [33]. (**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.

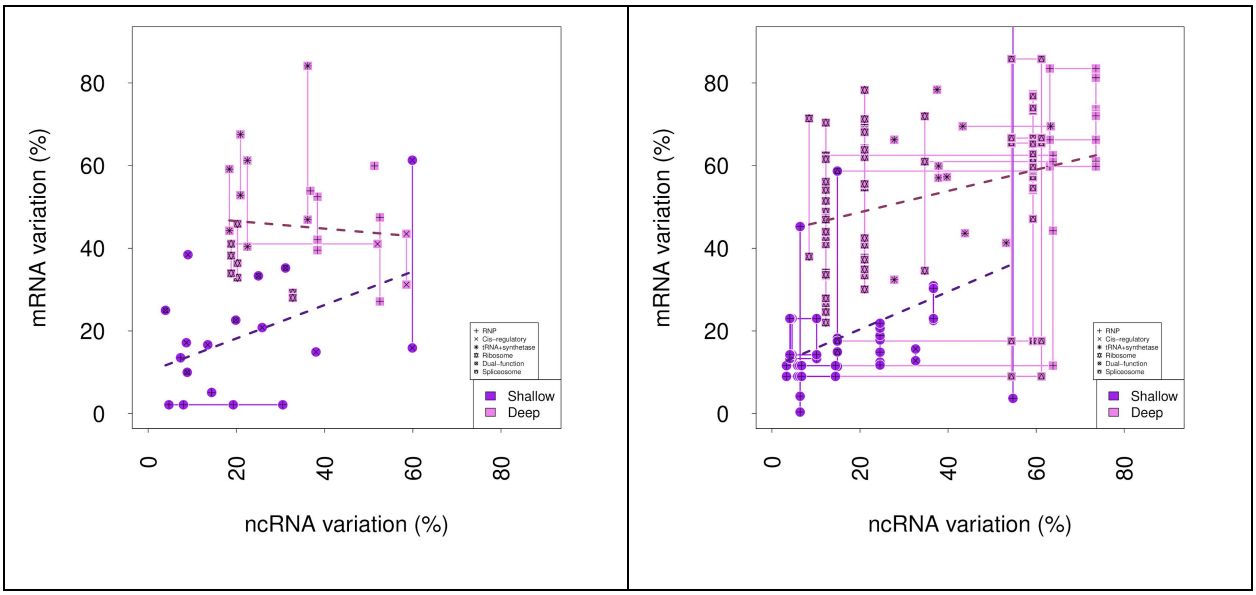


Figure S2: Level of nucleotide variation interacting RNA and protein pairs. Interacting RNA and protein pairs were compared across shallow and deep timescales in bacteria (left) and fungi (right). The percent nucleotide variation each RNA-protein pair was determined, for either shallow or deep divergence times. There was little correlation (Spearman's) between the variation in one molecule and its interacting partner in either the bacteria deep (*Rho* = -0.04, $p = 0.85$) or shallow (*Rho* = 0.29, $p = 0.24$) or fungi deep (*Rho* = 0.37, $p = 0.0001$) or shallow (*Rho* = 0.26 , $p = 0.07$) groups.

Maintenance of structure after in *silico* mutagenesis of RNPs

Table S6: Columns 1-6 contain the proportions of simulated point mutations in the protein and RNA components of RNPs from the PDB that can be classified as destabilising, divergent or both using estimated $\Delta\Delta G$ and Δbitscore values (see Figure 2A&B). Columns 7 and 8 contain the Spearman correlation coefficients and corresponding P-values for the relationships between the protein and RNA $\Delta\Delta G$ and Δbitscore values (see also Figure S3). In each case there is a significant relationship between the two, except for proteins with a single variant site.

	Destabilising ($\Delta\Delta G>0$)		**Divergent (Δbitscore<0)**		**Destabilising & Divergent ($\Delta\Delta G>0$ & Δ bitscore<0)**		**Spearman correlation coefficient (P-value)**	
	Protein	**RNA**	**Protein**	**RNA**	**Protein**	**RNA**	**Protein**	**RNA**
One mutn	30%	41%	50%	61%	60%	66%	-0.01 (0.67)	-0.35 ($<2.2e^{-16}$)
Four mutns	58%	73%	89%	96%	98%	93%	-0.20 ($<2.2e^{-16}$)	-0.25 ($<2.2e^{-16}$)

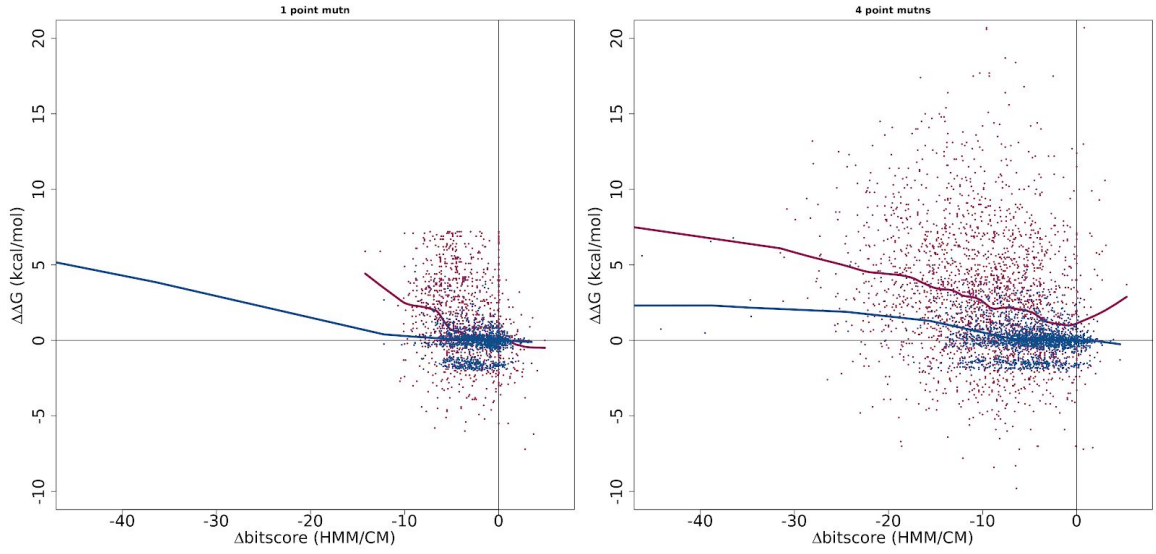


Figure S3: Relationship between $\Delta\Delta G$ and Δbitscore values for simulated protein and RNA mutations. Random mutations were introduced into mRNAs (blue) or ncRNAs (pink) for RNPs with solved structures. $\Delta\Delta G$ and Δbitscore values were computed as described in the methods, and have been used as an estimate of the impact of random simulated mutations on protein and RNA structures.

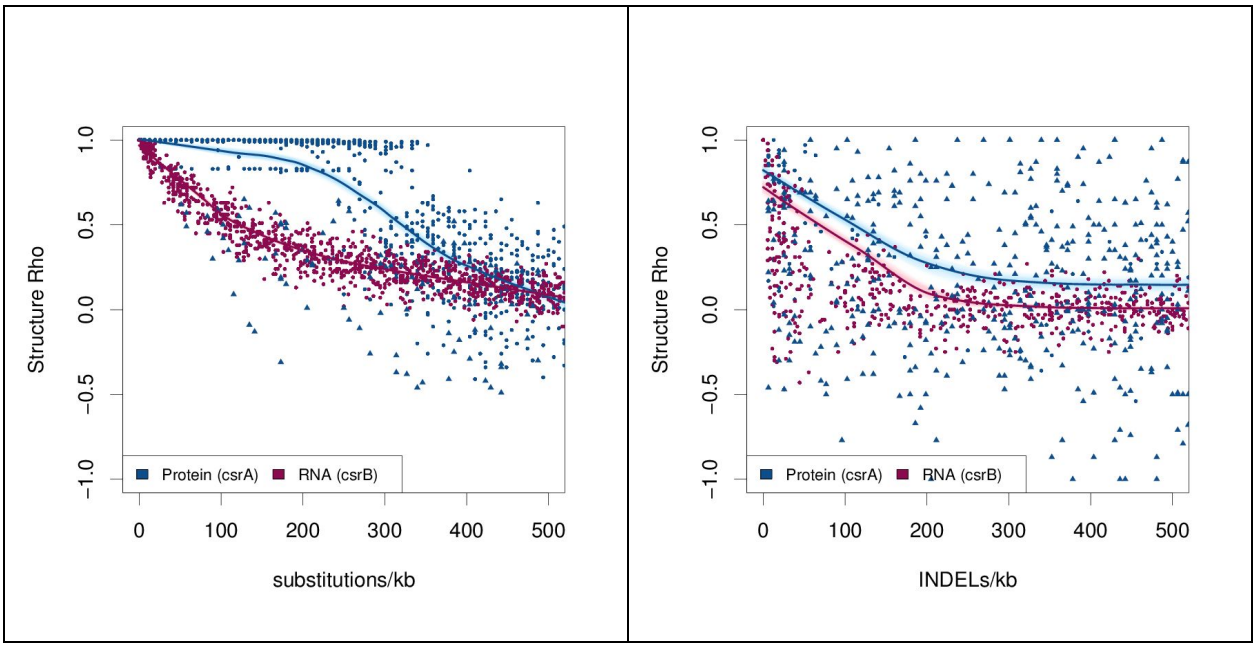


Figure S4: Structural robustness of the CsrA protein and CsrB sRNA. Random mutants of the CsrA messenger RNA (blue) and CsrB small RNA (pink) were generated *in silico*. Their secondary structure probabilities were predicted using "RNAfold-p" and "PSSpred". The per-residue probabilities of either base-paired/not-base-paired 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 S5: Parental plasmids for whole-plasmid PCR mutagenesis of mCherry and Broccoli

https://www.addgene.org/browse/sequence/114437/

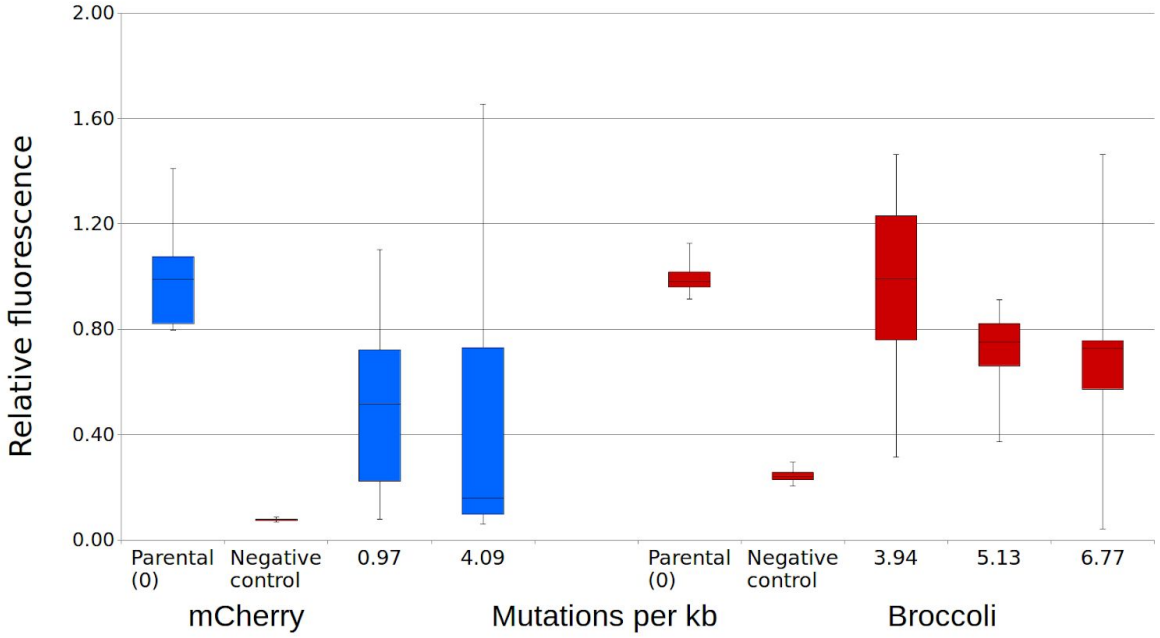


Figure S6: Fluorescence of mutant libraries of RNA aptamer Broccoli and protein mCherry. Libraries of randomly mutated fluorescent RNA aptamer Broccoli and fluorescent protein mCherry were tested for function relative to the unmutated control. Two libraries of mCherry and three libraries of Broccoli were constructed, 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. Waters LS, Storz G. Regulatory RNAs in bacteria. Cell. 2009;136: 615–628.

2. Cech TR, Steitz JA. The noncoding RNA revolution-trashing old rules to forge new ones. Cell. 2014;157: 77–94.

3. Gilbert W. Origin of life: The RNA world. Nature. 1986;319.

4. Lindgreen S, Umu SU, Lai AS-W, Eldai H, Liu W, McGimpsey S, et al. Robust identification of noncoding RNA from transcriptomes requires phylogenetically-informed sampling. PLoS Comput Biol. 2014;10: e1003907.

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

6. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceedings of the National Academy of Sciences. 2009;106: 9362–9367.

7. Chen J-Q, Wu Y, Yang H, Bergelson J, Kreitman M, Tian D. Variation in the ratio of nucleotide substitution and indel rates across genomes in mammals and bacteria. Mol Biol Evol. 2009;26: 1523–1531.

8. de Visser JAGM, Hermisson J, Wagner GP, Ancel Meyers L, Bagheri-Chaichian H, Blanchard JL, et al. Perspective: Evolution and detection of genetic robustness. Evolution. 2003;57: 1959–1972.

9. van Nimwegen E, Crutchfield JP, Huynen M. Neutral evolution of mutational robustness. Proceedings of the National Academy of Sciences. 1999;96: 9716–9720.

10. Fontana W, Stadler PF, Bornberg-Bauer EG, Griesmacher T, Hofacker IL, Tacker M, et al. RNA folding and combinatory landscapes. Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics. 1993;47: 2083–2099.

11. Kimura M. The Neutral Theory of Molecular Evolution. Cambridge University Press; 1984.

12. Omer S, Harlow TJ, Gogarten JP. Does Sequence Conservation Provide Evidence for Biological Function? Trends Microbiol. 2017;25: 11–18.

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

14. Sanjuan R. In Silico Predicted Robustness of Viroids RNA Secondary Structures. I. The Effect of Single Mutations. Mol Biol Evol. 2006;23: 1427–1436.

15. Mimouni NK, Lyngsø RB, Griffiths-Jones S, Hein J. An analysis of structural influences on

selection in RNA genes. Mol Biol Evol. 2009;26: 209–216.

16. Li C, Qian W, Maclean CJ, Zhang J. The fitness landscape of a tRNA gene. Science. 2016;352: 837–840.

17. Stombaugh J, Zirbel CL, Westhof E, Leontis NB. Frequency and isostericity of RNA base pairs. Nucleic Acids Res. 2009;37: 2294–2312.

18. Abrusán G, Marsh JA. Alpha Helices Are More Robust to Mutations than Beta Strands. PLoS Comput Biol. 2016;12: e1005242.

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

20. Goldberg AL, Wittes RE. Genetic Code: Aspects of Organization. Science. 1966;153: 420–424.

21. Alkatib S, Scharff LB, Rogalski M, Fleischmann TT, Matthes A, Seeger S, et al. The contributions of wobbling and superwobbling to the reading of the genetic code. PLoS Genet. 2012;8: e1003076.

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

23. Alff-Steinberger C. The genetic code and error transmission. Proc Natl Acad Sci U S A. 1969;64: 584–591.

24. Haig D, Hurst LD. A quantitative measure of error minimization in the genetic code. J Mol Evol. 1991;33: 412–417.

25. Novozhilov AS, Wolf YI, Koonin EV. Evolution of the genetic code: partial optimization of a random code for robustness to translation error in a rugged fitness landscape. Biol Direct. 2007;2: 24.

26. Wagner A. Robustness and Evolvability in Living Systems. 2013.

27. Geyer R, Madany Mamlouk A. On the efficiency of the genetic code after frameshift mutations. PeerJ. 2018;6: e4825.

28. Bartonek L, Braun D, Zagrovic B. Invariants of Frameshifted Variants. bioRxiv. 2019. p. 684076. doi:10.1101/684076

29. Maquat LE. Defects in RNA splicing and the consequence of shortened translational reading frames. Am J Hum Genet. 1996;59: 279–286.

30. MacArthur DG, Manolio TA, Dimmock DP, Rehm HL, Shendure J, Abecasis GR, et al. Guidelines for investigating causality of sequence variants in human disease. Nature.

2014;508: 469–476.

31. Ohta T. Slightly deleterious mutant substitutions in evolution. Nature. 1973;246: 96–98.

32. Chiu DK, Kolodziejczak T. Inferring consensus structure from nucleic acid sequences. Comput Appl Biosci. 1991;7: 347–352.

33. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A. 1992;89: 10915–10919.

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

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

36. Ferrada E, Wagner A. A Comparison of Genotype-Phenotype Maps for RNA and Proteins. Biophys J. 2012;102: 1916–1925.

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

38. Ahnert SE. Structural properties of genotype-phenotype maps. J R Soc Interface. 2017;14. doi:10.1098/rsif.2017.0275

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

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

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

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

43. Tsien RY. The green fluorescent protein. Annu Rev Biochem. 1998;67: 509–544.

44. Felsenstein J. DNAML in PHYLIP 2.6. University of Washington, Seattle. 1984.

45. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 2016;44:

D279–85.

46. Kalvari I, Argasinska J, Quinones-Olvera N, Nawrocki EP, Rivas E, Eddy SR, et al. Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. Nucleic Acids Res. 2018;46: D335–D342.

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

48. Jose BR, Gardner PP, Barquist L. Transcriptional noise and exaptation as sources for bacterial sRNAs. Biochem Soc Trans. 2019;47: 527–539.

49. Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW. Evolutionary rate in the protein interaction network. Science. 2002;296: 750–752.

50. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res. 2000;28: 235–242.

51. Tokuriki N, Stricher F, Schymkowitz J, Serrano L, Tawfik DS. The stability effects of protein mutations appear to be universally distributed. J Mol Biol. 2007;369: 1318–1332.

52. Ritz J, Martin JS, Laederach A. Evaluating our ability to predict the structural disruption of RNA by SNPs. BMC Genomics. 2012;13 Suppl 4: S6.

53. Wheeler NE, Barquist L, Kingsley RA, Gardner PP. A profile-based method for identifying functional divergence of orthologous genes in bacterial genomes. Bioinformatics. 2016;32: 3566–3574.

54. Wheeler NE, Gardner PP, Barquist L. Machine learning identifies signatures of host adaptation in the bacterial pathogen Salmonella enterica. PLoS Genet. 2018;14: e1007333.

55. Halvorsen M, Martin JS, Broadaway S, Laederach A. Disease-Associated Mutations That Alter the RNA Structural Ensemble. PLoS Genet. 2010;6: e1001074+.

56. Mattick JS, Makunin IV. Non-coding RNA. Hum Mol Genet. 2006;15 Spec No 1: R17–29.

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

58. Rost B. Twilight zone of protein sequence alignments. Protein Eng. 1999;12: 85–94.

59. Freyhult EK, Bollback JP, Gardner PP. Exploring genomic dark matter: a critical assessment of the performance of homology search methods on noncoding RNA. Genome Res. 2007;17: 117–125.

60. Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. Why highly expressed proteins evolve slowly. Proc Natl Acad Sci U S A. 2005;102: 14338–14343.

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

62. Eddy SR, Durbin R. RNA sequence analysis using covariance models. Nucleic Acids Res. 1994;22: 2079–2088.

63. Paige JS, Wu KY, Jaffrey SR. RNA mimics of green fluorescent protein. Science. 2011;333: 642–646.

64. Autour A, Westhof E, Ryckelynck M. iSpinach: a fluorogenic RNA aptamer optimized for in vitro applications. Nucleic Acids Res. 2016;44: 2491–2500.

65. Dolgosheina EV, Jeng SCY, Panchapakesan SSS, Cojocaru R, Chen PSK, Wilson PD, et al. RNA mango aptamer-fluorophore: a bright, high-affinity complex for RNA labeling and tracking. ACS Chem Biol. 2014;9: 2412–2420.

66. Introduction to Fluorescent Proteins. In: Nikon's MicroscopyU [Internet]. [cited 10 Aug 2018]. Available: https://www.microscopyu.com/techniques/fluorescence/introduction-to-fluorescent-proteins

67. Ellington AD, Szostak JW. In vitro selection of RNA molecules that bind specific ligands. Nature. 1990;346: 818–822.

68. Arnold FH. Design by Directed Evolution. Acc Chem Res. 1998;31: 125–131.

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

70. Lenski RE. Experimental evolution and the dynamics of adaptation and genome evolution in microbial populations. ISME J. 2017;11: 2181–2194.

71. Fontana W, Schuster P. Continuity in evolution: on the nature of transitions. Science. 1998;280: 1451–1455.

72. Gardner PP, Holland BR, Moulton V, Hendy M, Penny D. Optimal alphabets for an RNA world. Proc Biol Sci. 2003;270: 1177–1182.

73. Moulton V, Gardner PP, Pointon RF, Creamer LK, Jameson GB, Penny D. RNA folding argues against a hot-start origin of life. J Mol Evol. 2000;51: 416–421.

74. Lepper CP, Williams MAK, Edwards PJB, Filichev VV, Jameson GB. Effects of Pressure and pH on the Physical Stability of an I-Motif DNA Structure. ChemPhysChem. 2019. pp. 1567–1571. doi:10.1002/cphc.201900145

75. McBride RC, Ogbunugafor CB, Turner PE. Robustness promotes evolvability of

thermotolerance in an RNA virus. BMC Evol Biol. 2008;8: 231.

76. Lenski RE, Barrick JE, Ofria C. Balancing robustness and evolvability. PLoS Biol. 2006;4: e428.

77. Bloom JD, Labthavikul ST, Otey CR, Arnold FH. Protein stability promotes evolvability. Proc Natl Acad Sci U S A. 2006;103: 5869–5874.

78. Poole AM, Jeffares DC, Penny D. The path from the RNA world. J Mol Evol. 1998;46: 1–17.

79. Madera M, Gough J. A comparison of profile hidden Markov model procedures for remote homology detection. Nucleic Acids Res. 2002;30: 4321–4328.

80. Li Z, Zhang Y. Predicting the secondary structures and tertiary interactions of 211 group I introns in IE subgroup. Nucleic Acids Res. 2005;33: 2118–2128.

81. Lowe TM, Chan PP. tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. Nucleic Acids Res. 2016;44: W54–7.

82. Kalvari I, Argasinska J, Quinones-Olvera N, Nawrocki EP, Rivas E, Eddy SR, et al. Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. Nucleic Acids Res. 2018;46: D335–D342.

83. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics. 2013;29: 2933–2935.

84. Eddy SR. Accelerated Profile HMM Searches. PLoS Comput Biol. 2011;7: e1002195.

85. Suyama M, Torrents D, Bork P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 2006;34: W609–12.

86. Mann HB, Whitney DR. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. Ann Math Stat. 1947;18: 50–60.

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

88. Felsenstein J. PHYLIP version 3.6. Software package, Department of Genome Sciences, University of Washington, Seattle, USA. 2005.

89. Lorenz R, Bernhart SH, Höner Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, et al. ViennaRNA Package 2.0. Algorithms Mol Biol. 2011;6: 26.

90. Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, et al. eggNOG

4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res. 2016;44: D286–93.

91. Laimer J, Hofer H, Fritz M, Wegenkittl S, Lackner P. MAESTRO--multi agent stability prediction upon point mutations. BMC Bioinformatics. 2015;16: 116.

92. Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P. Fast folding and comparison of RNA secondary structures. Monatshefte für Chemie / Chemical Monthly. 1994;125: 167–188.

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

94. Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. The I-TASSER Suite: protein structure and function prediction. Nat Methods. 2015;12: 7–8.

95. McCaskill JS. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers. 1990;29: 1105–1119.

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