Avoidance of stochastic RNA interactions can be harnessed to control protein expression levels in bacteria and archaea

Authors:

Sinan U. Umu[1,2], Anthony M. Poole[1,2,3], Renwick C.J. Dobson[1,2,5], Paul P. Gardner[1,2,4]*

[1]School of Biological Sciences, [2]Biomolecular Interaction Centre, [3]Allan Wilson Centre for Molecular Ecology & Evolution, [4]Bio-Protection Research Centre, University of Canterbury, Christchurch, New Zealand

[5]Department of Biochemistry and Molecular Biology, University of Melbourne, Parkville, VIC 3010, Australia

*paul.gardner@canterbury.ac.nz

Abstract

A critical assumption of gene expression analysis is that mRNA abundances broadly correlate with protein abundance, but these two are often imperfectly correlated. Some of the discrepancy can be accounted for by two important mRNA features: codon usage and mRNA secondary structure. We present a new global factor, called mRNA:ncRNA avoidance, and provide evidence that avoidance increases translational efficiency. We also demonstrate a strong selection for avoidance of stochastic mRNA:ncRNA interactions across prokaryotes, and that these have a greater impact on protein abundance than mRNA structure or codon usage. By generating synonymously variant green fluorescent protein (GFP) mRNAs with different potential for mRNA:ncRNA interactions, we demonstrate that GFP levels correlate well with interaction avoidance. Therefore, taking stochastic mRNA:ncRNA interactions into account enables precise modulation of protein abundance.

Main Text

Introduction

It should in principle be possible to predict protein abundance from genomic data. However,

protein and mRNA levels are often poorly correlated (de Sousa Abreu et al. 2009; Vogel & Marcotte 2012; Kwon et al. 2014; Maier et al. 2011; Lu et al. 2007; Taniguchi et al. 2010; Chen et al. 2016), which is a major barrier to precision bioengineering and quantification of protein levels. mRNA secondary structure (Pelletier & Sonenberg 1987; Chamary & Hurst 2005), codon usage (Ikemura 1981; Andersson & Kurland 1990; Sharp & Li 1987), and mRNA (and protein) degradation rates (Maier et al. 2011) are commonly invoked to explain this discrepancy. Yet, at best, these features account for only 40% of variation, and in some instances explain very little of the observed variation (Kudla et al. 2009; Maier et al. 2011; Plotkin & Kudla 2011; Goodman et al. 2013; Chen et al. 2016). Here we show that crosstalk interactions between ncRNAs and mRNAs also impact protein abundance, and that such interactions have a greater effect than either mRNA secondary structure or codon usage. We measured interactions between a set of evolutionarily conserved core mRNAs and ncRNAs from 1,700 prokaryotic genomes using minimum free energy (MFE) models. For 97% of species, we find a reduced capacity for interaction between native RNAs relative to controls. Furthermore, by generating synonymously variant GFP mRNAs that differ in their potential to interact with core ncRNAs, we demonstrate that GFP expression levels can be both predicted and controlled. Our results demonstrate that there is strong selection for avoidance of stochastic mRNA:ncRNA interactions across prokaryotes. Applying this knowledge to mRNA design will enable precise control of protein abundance through the incorporation or exclusion of inhibitory interactions with native ncRNAs.

Result and Discussion

To examine if avoidance of stochastic mRNA:ncRNA interactions is a feature of transcriptomes in bacteria and archaea, we estimated the strength of all possible intermolecular RNA interactions using a minimum free energy (MFE) model (Mückstein et al. 2006). If stochastic interactions are selected against, because of the capacity for abundant ncRNAs (Lindgreen et al. 2014; Deutscher 2006; Giannoukos et al. 2012) to impact translation (Waters & Storz 2009; Storz et al. 2011), such negative selection would be most comparable between species and readily detected for broadly conserved ncRNAs and mRNAs. We computed the free energy distribution of interactions between highly conserved mRNA:ncRNA pairs and compared this to

a number of negative control interactions, which serve to show the background distribution of binding energy values (Figure 1A). The initiation of translation has been shown to be the rate limiting step for translation (Tuller & Zur 2015; Plotkin & Kudla 2011; Nakahigashi et al. 2014), therefore, we focus our analysis on the first 21 nucleotides of the mRNA coding sequence (CDS). This has the further advantage of reducing computational complexity. We also test a variety of negative control mRNA regions, which are unlikely to play a functional role in RNA:ncRNA interactions. The mRNA controls include (1) di-nucleotide preserving shuffled sequences (Workman & Krogh 1999) (orange, Figure 1A), (2) homologous mRNAs from another phylum (with compatible guanine-cytosine (G+C) content) (purple), (3) downstream regions 100 base pairs (bps) within the CDS (pink), (4) the reverse complement of the 5` of CDSs (green), and lastly (5) unannotated (intergenic) genomic regions (yellow). Our interaction predictions in a single model strain show that native interactions consistently have higher (i.e. less stable) free energies than expected when compared to the five different mRNA negative controls: that is, there is a reduced capacity for native mRNAs and native ncRNAs to interact. We also compared different energy models and confirm that the MFE shift is a result of intermolecular binding (Figure 1-figure supplement 1A,B,C). We subsequently deployed the most conservative negative control (di-nucleotide preserving shuffle) and free energy model (Figure 1-figure supplement 1C) to detect if this shift for less stable binding of mRNA:ncRNA is true of all bacteria and archaea.

In terms of stoichiometry, the model we use assumes that ncRNA expression levels are vastly in excess of mRNA expression levels (i.e. [ncRNA]>>[mRNA]) (Giannoukos et al. 2012; Deutscher 2006). This is generally a biologically reasonable assumption when focussing on core genes based upon past analysis and our own work with RNA-seq data from a range of bacteria and archaea (Figure 4) (Lindgreen et al. 2014). Consequently, any potential mRNA interaction regions are saturated with ncRNA, therefore a summative model of interaction energies is a reasonable approximation to the estimated impact of excess hybridisation. If modelling ncRNAs that are not so abundant, then a model weighted by expression level may be advantageous, but it is difficult to determine these across multiple conditions and developmental stages.

In order to assess whether mRNA:ncRNA is an evolutionarily conserved phenomenon, we calculated intermolecular binding energies for conserved ncRNAs and mRNAs from 1,582 bacterial and 118 archaeal genomes and compared these to a negative control dataset derived using a dinucleotide frequency preserving shuffling procedure (Workman & Krogh 1999). This measures a property that we call the `extrinsic avoidance` of mRNA:ncRNA interactions, yet this approach may fail to identify genuine avoidance in cases when the G+C content differences between interacting RNAs is extreme. Measuring only extrinsic avoidance (using shuffled mRNAs as negative controls), we found that stochastic mRNA:ncRNA interactions are significantly underrepresented in most (73%) of the prokaryotic phyla ($P < 0.05$, one-tailed Mann-Whitney U test) (Figure 1B,C and Figure 1-figure supplement 2). This indicates that there is selection against stochastic interactions in both bacteria and archaea.

We next sought to establish the degree to which intrinsic G+C features of RNAs lead to avoidance of stochastic interactions (Figure 1D). A test of G+C composition revealed a significant difference ($P < 0.05$, two-tailed Mann-Whitney U test) between mRNAs and ncRNAs for 95% of bacteria and archaea (Figure 1D,E). Therefore, either extrinsic or intrinsic avoidance signals indicate that selection against stochastic interactions and it is near-universal for the prokaryotes (97% of all strains) (Figure 1E and Supplementary file 1A,B).

Our results clearly establish a signature of selection that acts to minimise stochastic mRNA:ncRNA interactions. However, with thousands of potential interacting RNA species in even simple prokaryotic systems (Vivancos et al. 2010; Sharma et al. 2010), the complete avoidance of stochastic interactions is combinatorially unlikely. Therefore there ought to be a tradeoff between avoidance and optimal expression. To assess this, we examined the relationship between potential stochastic interactions and the variation between mRNA and cognate protein levels for four previously published endogenous mass spectrometry datasets from *Escherichia coli (E. coli)* and *Pseudomonas aeruginosa* (Laurent et al. 2010; Kwon et al. 2014; Lu et al. 2007). We computed Spearman's correlation coefficients between protein abundances and

extrinsic avoidance, 5` end internal mRNA secondary structure and codon usage. Of the three measures, avoidance is significantly correlated in all four datasets (Spearman's rho values are between 0.11 - 0.17 and corresponding *P* values are between 0.01 - $1.3x10^{-12}$) (Supplementary file 3). In contrast, 5` end mRNA structure significantly correlates in two datasets, and codon usage significantly correlates in all four datasets. This indicates that, despite strong selection against stochastic interactions, such interactions do significantly impact the proteome (Figure 2A).

We also test how mRNA:ncRNA crosstalk impacts the translation of transformed mRNAs that have not coevolved with the ncRNA repertoire. We examined two available *E. coli*-based GFP experimental datasets (Goodman et al. 2013; Kudla et al. 2009), where synonymous mRNAs are generated for a GFP reporter gene. This enables the assessment of the impact of synonymous changes on protein abundance using fluorescence. Avoidance and mRNA secondary structure are both significantly correlated with fluorescence, whereas codon usage is not (Spearman's rho values are 0.11 and 0.65 , the corresponding *P* values are $3.17x10^{-41}$ and 1.69^{-20}) (Figure 2A). Note that one of the GFP datasets (Goodman et al. 2013) uses native *E. coli* mRNA 5` ends for their constructs, whereas the other GFP dataset (Kudla et al. 2009) is randomly generated. We observe that the influence of avoidance on gene expression for randomly sampled synonymous mRNAs is strong, while endogeneous gene expression is limited. Presumably, due to negative selection pruning low avoidance mRNAs from the gene pool (Figure 2A).

Our results indicate that crosstalk between mRNAs and ncRNAs can impact protein expression levels. We therefore predict that taking crosstalk into account will enable the design of constructs where protein expression levels can be precisely controlled. To test this, we generated GFP constructs based on the following constraints: codon bias, 5` end mRNA secondary structure stability and crosstalk avoidance (see 'Materials and methods'). Our constructs are designed to capture the extremes of one variable, while controlling other variables (e.g. high or low avoidance and near-average codon bias and mRNA secondary structure). The G+C content, a known confounding factor, was also strictly controlled for each construct. We selected a

commercial service to perform our GFP transformations to avoid possible bias and increased the robustness of our approach (Ioannidis & Khoury 2011). We predicted that a construct where all three parameters are optimised will result in a high expression. Consistent with predictions, our optimised construct had maximal expression (Figure 2-figure supplement 1). Of the three parameters, avoidance showed the largest range, suggesting that tuning this parameter permits expression levels to be finely controlled ($R_s = 0.56$, $P = 6.9x10^{-6}$) (Figure 2B,C,D and Figure 2-figure supplement 2A,B,C).

For a final confirmation of the avoidance hypothesis, we tested the *Thermus thermophilus SSU* ribosomal RNA, which is a component of one of the most complete prokaryotic ribosomal structures available in the PDB (Rozov et al. 2015). We identified the regions of the SSU rRNA that had the least capacity to interact with *T. thermophilus* core mRNAs and found that these regions were generally not bound to either ribosomal proteins or other ncRNAs, such as the LSU rRNA ($P = 2.49x10^{-17}$, Fisher's exact test) (Figure 3; see 'Materials and methods').

This study focusses on the 5` ends of the CDS as this region is important for the initiation of translation (Plotkin & Kudla 2011; Tuller & Zur 2015) and is a consistent feature of all the genomic, transcriptomics, proteomics and GFP expression datasets that we have evaluated in this work. In smaller-scale tests we have observed similar conserved avoidance signals within the entire CDSs (Figure 3-figure supplement 1) and within the 5`UTRs (Figure 3-figure supplement 2). Furthermore, we predict that similar signals can be observed for mRNA:mRNA and ncRNA:ncRNA avoidance. Although the impacts of these features are challenging to validate, interactions between clustered regularly interspaced short palindromic repeats (CRISPR) spacer sequences (Bhaya et al. 2011) and core ncRNAs are good candidates to test ncRNA:ncRNA avoidance.

In conclusion, our results indicate that the specificity of prokaryotic ncRNAs for target mRNAs is the result of selection both for a functional interaction and against stochastic interactions. Our experimental results support the view that stochastic interactions are selected against, due to

deleterious outcomes on expression. We suspect avoidance of crosstalk interactions has several evolutionary consequences. First, as transcriptional outputs become more diverse in evolution, we expect that the probability of stochastic interactions for both new ncRNAs and mRNAs becomes higher. This will impact the emergence of new, high abundance RNAs, since selection for high abundance may be mitigated by deleterious crosstalk events. Second, we predict that stochastic interactions limit the number of simultaneously transcribed RNAs, since the combinatorics of RNA:RNA interactions imply that eventually stochastic interactions cannot be avoided. This may in turn drive selection for forms of spatial or temporal segregation of transcripts. Finally, taking codon usage, mRNA secondary structure and potential mRNA:ncRNA interactions into account allows better prediction of proteome outputs from genomic data, and informs the precise control of protein levels via manipulation of synonymous mRNA sequences (Figure 2-figure supplement 3).

Materials and Methods

Here we summarize the data sources, materials and methods corresponding to our manuscript. We performed all statistical analyses in R, and all other computational methods in Python 2.7 or Bash shell scripts. We explicitly cite all the bioinformatics tools and their versions.
All tables (Supplementary files 1-4) are available as supporting online material. All of our own sequences, scripts and R workspace images are available on Github including the supplementary files. (github.com/UCanCompBio/Avoidance). The other datasets are cited in the manuscript (Supplementary file 3).

Evolutionary conservation

If excessive interactions between messenger RNAs (mRNAs) and non-coding RNAs (ncRNAs) are detrimental to cellular function, then we expect the signature of selection against interactions (avoidance) to be a conserved feature of prokaryotic genomes. In the following we describe where the data used to to test the evolutionary conservation of avoidance was acquisitioned, the models that we use to test avoidance and the negative controls in detail for evolutionary conservation predictions. We also investigate detect regions of avoidance on one of the core

ncRNAs, the ribosomal small subunit (SSU) RNA.

Data sources for bacterial genomes

The bacterial genomes and annotations that we used for investigating mRNA:ncRNA interactions were acquired from the EBI nucleotide archive (2,564 sequenced bacterial genomes available on August 2013) (http://www.ebi.ac.uk/genomes/bacteria.html). We selected an evolutionarily conserved (core) group of 114 mRNAs from PhyEco (Wu et al. 2013) and an evolutionarily conserved (core) group of ncRNAs(Hoeppner et al. 2012). PhyEco markers are based on a set of profile HMMs that correspond to highly conserved bacterial protein coding genes (these include ribosomal proteins, tRNA synthetases as well as other components of translation machinery, DNA repair and polymerases) (Wu et al. 2013). The HMMer package (version 3.1b1) (Eddy 2011) was used to extract the mRNAs corresponding to these marker genes from genome files. We removed genome sequences that host fewer than 90% of the marker genes; leaving 1,582 bacterial genome sequences and 176,704 core mRNAs that spanned these.

We extracted the 1st to the 21st nucleotide of the core mRNAs. As this region showed the strongest signal in a small-scale analysis (Figure 3-figure supplement 1A), this region has also been shown to have an unusual codon distribution in previous work (Tuller & Zur 2015; Goodman et al. 2013) as explained in the main manuscript.

We obtained ncRNA annotations using the Rfam database (version 11.0)(Gardner et al. 2011) for the well conserved and highly expressed tRNA, rRNA, RNase P RNA, SRP RNA, tmRNA and 6S RNA families (Rfam accessions: RF00001, RF00005, RF00010, RF00011, RF00013, RF00023, RF00169, RF00177). The redundant annotations were filtered for overlapping and identical paralogous sequences, leaving 99,281 core ncRNA that spanned 1,582 bacterial genomes.

Data sources for archaeal genomes

We followed a similar pipeline for archaeal genomes as described for bacterial genomes. In total we processed 240 archaeal genomes, and after filtering those that had fewer than 90% of the

marker genes, we had 118 archaeal genomes for further analysis (genomes available on August 2013) (http://www.ebi.ac.uk/genomes/archaea.html). These genomes host 12,370 and 10,804 core mRNAs and core ncRNAs respectively.

Test of an (extrinsic) avoidance model

We used RNAup (version 2.0.7)(Lorenz et al. 2011) to calculate the binding minimum (Gibbs) free energy (MFE) values of mRNA:ncRNA interactions. The RNAup algorithm combines the intramolecular energy necessary to open binding sites with intermolecular energy gained from hybridization (Mückstein et al. 2006). In other words, this approach minimizes the sum of opening intramolecular energies and the intermolecular energy (Figure 1-figure supplement 1C). In our model of avoidance, we test for a reduction in absolute binding MFE relative to negative controls as a measure of avoidance.

After testing a variety of negative controls (e.g. dinucleotide preserved shuffled mRNAs, the 5` end of homologous mRNAs from a different bacterial phylum, 100 nucleotides downstream of designated interaction region, reverse complements, and identically sized intergenic regions), we selected the dinucleotide frequency preserved shuffled sequences as our negative control since this displayed the most conservative interaction MFE distribution (Figure 3-figure supplement 1A,B,C). In more detail, to serve as a negative control we compute the interaction MFE between each of the core ncRNAs and 200 dinucleotide-preserved shuffled versions of the 5` end mRNAs. A dinucleotide frequency preserving shuffling procedure is used as Gibbs free energies are computed over base pair stacks, i.e. a dinucleotide alphabet, therefore this method has been shown to be important in order to minimise incorrect conclusions (Workman & Krogh 1999).

We tested if the energy difference between native and shuffled interaction distributions is statistically significant using the nonparametric one-tailed Mann–Whitney U test, which returns a single *P* value per genome (Figure 1C). If the distribution of native interaction energies for a genome is significantly higher (i.e. fewer stable interactions) than the negative control, this is an indication that the genome has undergone selection for mRNA:ncRNA avoidance. To create the background density difference lines (seen in grey at Figure 1B), we randomly selected 100 bacterial strains and plot differences between the densities of shuffled interactions.

Test of an intrinsic avoidance model

The energy-based avoidance model that we defined above is opaque to cases of "intrinsic avoidance". These are where the intrinsic properties of mRNA and ncRNA sequences restrict their ability to interact. For an extreme example, if ncRNAs are composed entirely of guanine and cytosine nucleotides, whilst mRNAs are composed entirely of adenine and uracil nucleotides, then these will rarely interact. Therefore, our energy-based avoidance measures for native and shuffled interactions will both be near zero, and thus will not detect a significant energy shift between the native and control sequences. In order to account for some of these issues, we compared the G+C difference between core ncRNAs and core mRNAs. We used a nonparametric two-tailed Mann-Whitney U test to determine if there is a statistically significant G+C difference between the two samples: G+C of ncRNAs vs G+C of 5` end mRNAs (Figure 1D, E).

Sliding window analysis to detect regions of significance for avoidance on SSU ribosomal RNA

We hypothesise that heterogeneous signals of avoidance within ncRNA sequences may correspond to the accessibility of different ncRNA regions. For example, are highly avoided regions of abundant ncRNAs more accessible than those that are avoided less? To create an avoidance profile, we tested binding MFEs of native and shuffled interactions throughout the full-length SSU ribosomal RNA of *Thermus thermophilus* HB8 (*T. thermophilus*), using a one tailed Mann-Whitney U-tests to evaluate the degree of avoidance for each nucleotide in the SSU rRNA (Figure 3) with a windows size of 10 and step size of 1 (Supplementary file 4). We selected the protein data bank (PDB) entry (4WZO) as it is one of the few ribosomal structures with associated protein, mRNA, tRNA and LSU binding data (Rozov et al. 2015). The native interactions are the interactions between *T. thermophilus* core mRNAs and SSU ribosomal RNA. The shuffled controls are derived from 200 dinucleotide preserved shuffled versions of the RNAs. We created a 2x2 contingency table which separates the counts of residues that either host a strong avoidance signal or little avoidance signal (regions with $P < 0.001$, Mann-Whitney U-test) and residues that we predict to either be in contact ($\leq$3.4 Angstroms between atoms) with

ribosomal proteins or ribosomal, transfer or messenger RNAs or not in contact with other molecules (i.e. accessible) (Figure 3). We applied a Fisher's exact test (Fisher 1922) to these groups to and discovered a statistically significant relationship between avoidance and accessibility ($P = 2.5\text{x}10^{-17}$).

Proteomics/Transcriptomics & GFP expression

We predict that mRNAs with low avoidance values will produce fewer proteins for each mRNA transcript than those with high avoidance. In order to test this, we conducted a meta-analysis of proteomics and transcriptomics data and the relationship between this data and measures of mRNA and ncRNA avoidance. In the following section we describe the origins of the data we have used and the statistical analysis we use to test whether avoidance influences gene expression.

Data sources and statistics for mRNA, protein abundance & GFP expression

We compiled our data from five protein and mRNA quantification datasets, which consist of three *Escherichia coli* (Laurent et al. 2010; Goodman et al. 2013; Lu et al. 2007) and two *Pseudomonas aeruginosa* (Laurent et al. 2010; Kwon et al. 2014). We calculated Spearman's correlation coefficients (and associated *P* values) among the protein abundances and 5` end secondary structure (measured by intermolecular MFE), codon bias (measured by codon adaptation index (CAI)) and avoidance (Figure 2A).

CAI metric defines how well mRNAs are optimised for codon bias (Sharp & Li 1987) . The CAI values were determined based on codon distribution patterns acquired from the core protein coding genes of *E. coli* BL21(DE3) (Accession: AM946981.2) (Wu et al. 2013) using Biopython libraries (version 1.6) (Cock et al. 2009).

The folding MFE predicts how stable the secondary structure of an RNA can be. The folding MFEs of GFP mRNAs were calculated using the RNAfold algorithm (version 2.0.7) (Lorenz et al. 2011). We restricted folding energy to first 37 nucleotides because the most significant correlation was previously reported for this region (Kudla et al. 2009).

We acquired previously published GFP data, associated fluorescence values and mRNA

quantifications (Kudla et al. 2009) via personal communication. Our avoidance model showed the highest and most significant correlation with GFP expression in that dataset (R_s = 0.65, P = $1.69x10^{-20}$) (Figure 2A and Figure 2-figure supplement 2D,E,F). 5` end secondary structure (R_s = 0.62, P = $5.73x10^{-18}$) correlates slightly less than avoidance, while CAI does not correlate significantly (R_s = 0.02, P = 0.4).

Sliding window analysis to detect regions of significance for avoidance on mRNAs

In order to identify a region of mRNA that is consistent and unique in the datasets that we applied evolutionary and expression analyses to we created an avoidance profile from the previously published GFP mRNAs (Kudla et al. 2009). We calculated binding MFEs using a window size of 21 with a 1 nucleotide step size, and for each region we computed the associated Spearman's correlation coefficients with P values. This analysis revealed the significance of the first 21 nucleotides on expression, this is consistent with previous results that identify initiation as the rate limiting step for translation (Tuller & Zur 2015; Plotkin & Kudla 2011). It also revealed other statistically significant regions with high correlation correlation coefficient throughout the GFP mRNAs (Figure 3-figure supplement 1A).

mRNA design

We have shown that avoidance is a broadly evolutionary conserved phenomenon and that it is significantly correlated with protein abundance relative to mRNA abundance. We now wish to test if avoidance can be used to design mRNA sequences that modulate the abundance of corresponding protein in a predictable fashion. We use a set of GFP mRNA constructs that all maintain the same G+C content, codon adaptation index (CAI) and internal secondary structure but host either very high or very low avoidance values. This procedure was repeated for the CAI and internal secondary structure values while maintaining a constant avoidance. The resulting 13 constructs were synthesised, transformed and expressed by commercial services. In the following paragraphs we explained how we design our GFP constructs, the experimental set-up and statistical analyses.

Green fluorescence protein (GFP) mRNA design

We sampled 537,000 synonymous mRNA variants of a GFP mRNA (the 239 AA, 720 nucleotide long, with accession AHK23750, can be encoded by 7.62×10^{111} possible unique mRNA variants). In brief, these mRNA variants were scored based upon (1) CAI, (2) mRNA secondary structure in their 5` end region, and (3) mRNA:ncRNA interaction avoidance in their 5` end region.

The genome of *E. coli* BL21 encodes 52 unique core ncRNAs (Gardner et al. 2011), to estimate the level of ncRNA avoidance for each GFP mRNA, we sum the binding MFEs. For example, for each GFP mRNA we compute 52 independent binding MFE values for each ncRNA. In short, a higher summed MFE score for a GFP mRNA implies a higher avoidance, while a lower summed MFE score implies a lower avoidance. This approach assumes that the ncRNAs are expressed at much higher levels than GFP mRNAs (i.e. [ncRNA] >> [mRNA]) (Figure 4). Consequently, any potential interaction site on GFP mRNAs are likely to be saturated with ncRNA.

Finally, we selected 13 GFP mRNA constructs, while controlling the range of G+C values. These GFP mRNAs were designed to have four different aspects; extreme 5` end secondary structure (2 minimum and 2 maximum folding MFE constructs), extreme codon bias (2 maximum and 2 minimum CAI constructs), extreme interaction avoidance (2 minimum and 2 maximum binding MFE constructs) and an "optimal" construct. The optimal construct was selected for a high CAI, low 5` end structure and high avoidance. All extreme GFP mRNA constructs have near identical G+C content (between 0.468-0.480) and identical G+C contents at the 5` end (0.48). Each of the sampled GFP mRNAs is separated from other mRNAs by at least 112 nucleotide substitutions and 122 nucleotide substitutions on average (Figure 2-figure supplement 1).

Extreme GFP transformations, determining fluorescence levels and RT-qPCR analyses

Both GFP expression assays and RT-qPCR analyses were performed as part of a commercial service offered by the University of Queensland, Protein Expression Facility and Real-Time PCR Facility.

Plasmid DNA from each construct was transformed into a expression strain of *E. coli* BL21(DE3). Starter cultures were grown in quadruplicate from single colonies in 0.5 mL of TB kanamycin 30 µg/mL media in a 96 deep-well microplate and incubated at 30ºC, 400 rpm (3 mm shaking throw). Each starter culture was used to inoculate 1.0 mL of the same media at a ratio of 1:50, each in a single well of a 96 deep-well plate. The cultures were incubated at 30°C, 400 rpm for 1 hour, at this point the cultures were chilled for 5 min then induced into 0.2 mM IPTG and incubated at 20°C. For analysis, culture samples of 100 µL were taken at 1 hr, 2 hrs, 3 hrs, 4 hrs and 22 hrs (overnight) hours post-induction (HPI) for fluorescence and optical density analysis. Samples were collected in PetriWell 96-well flat bottom, black upper, lidded microplates (Genetix). Cell density of fluorescence measurements were performed on a Spectramax M5 Microplate Reader using SMP software v 5.2 (Molecular Devices). For fluorescence intensity measurements, samples were collected in the 96-well plate listed above. Samples were analysed by bottom-read, 10 reads per well at an excitation wavelength = 488 nm, emission wavelength = 509 nm with an automatic cut-off at 495 nm and measured as relative fluorescence units (RFU). The raw RFU values were normalised by subtracting the averaged baseline values obtained from untransformed BL21(DE3) at the same time point. All samples at the 22 HPI time point were diluted 1:4 in TB kanamycin 30 µg/mL media before measurement.

Total RNA was purified from induced 0.5 mL of BL21(DE3) cultures on Maxwell® 16 robot (Promega) using LEV simplyRNA Tissue Kit (Promega). RNA concentrations were assessed on Qubit 3.0 Fluorometer (Thermo Fisher Scientific). cDNA synthesis was done using ProtoScript II First Strand cDNA Synthesis Kit (NEB) according to manufacturer protocol using random primer. The rpsL gene was selected as the reference gene (internal control). RT-qPCR was performed in 384-well plates with a ViiA™ 7 Real-Time PCR System (Thermo Fisher Scientific) using Life Technology SYBR Green-based PCR assay. The data analysis was performed using Applied Biosystems QuantStudio software (Thermo Fisher Scientific). The total volume of reaction was 10 µL including 0.2 µM of each primer as a final concentration. The following PCR conditions were used: 95ºC for 10 min, followed by 40 cycles of 95ºC for 15 s and 60ºC for 1 min. The melting curves were analyzed at 60-95ºC after 40 cycles. RNA concentrations were subsequently estimated using the $2^{-\Delta\Delta C_T}$ approach (Schmittgen & Livak

2008). We shared the raw data, oligos and primers in the supplementary files (Supplementary file 2A,B).

Statistical analyses of extreme GFP data

As described, we designed extreme GFP mRNA constructs, and measured the associated fluorescence. A Kruskal-Wallis test (nonparametric alternative of ANOVA) shows a statistically significant difference between the fluorescence of GFP mRNA groups ($P = 1.35\text{x}10^{-5}$) (Figure 2-figure supplement 1). Our pairwise comparison of GFP groups using a Kruskal-Nemenyi test (a nonparametric alternative of the Student's t-test) for fluorescence difference also reveals a statistically significant difference in fluorescence between high avoidance constructs and low avoidance constructs ($P = 0.00036$).

We computed the Spearman's correlation coefficients (and associated P values) between GFP expression and each of the following measures; CAI ($R_s = 0.29$, $P = 0.016$), intramolecular folding energy ($R_s = 0.34$, $P = 0.006$), avoidance (intermolecular binding energy) ($R_s = 0.56$, $P = 6.9\text{x}10^{-6}$) and mRNA concentration ($R_s = 0.73$, $P = 3.2\text{x}10^{-3}$) to predict effect size of each predictor. Our avoidance model resulted in the highest correlation with GFP expression (Figure 2B,C,D).

Acknowledgements

SUU is supported by a Biomolecular Interaction Centre and UC HPC (Bluefern) joint PhD Scholarship from the University of Canterbury. AMP & PPG are both supported by Rutherford Discovery Fellowships, administered by the Royal Society of New Zealand. RCJD acknowledges the Royal Society of New Zealand Marsden Fund and US Army Research Office for funding support. Thanks to Grzegorz Kudla for sharing the GFP expression data from Kudla *et al* (2009) and Cindy Chang, Emilyn Tan, Michael Nefedov from the RT-PCR and the PEF facilities at the University of Queensland for assistance with generating GFP expression data. We also acknowledge Jeppe Vinther, Lukasz Kielpinski, Anders Krogh and the attendees of the 2012 and 2015 Benasque RNA conference for stimulating discussions.

References

Andersson, S.G. & Kurland, C.G., 1990. Codon preferences in free-living microorganisms. *Microbiological reviews*, 54(2), pp.198–210.

Bhaya, D., Davison, M. & Barrangou, R., 2011. CRISPR-Cas systems in bacteria and archaea: versatile small RNAs for adaptive defense and regulation. *Annual review of genetics*, 45, pp.273–297.

Chamary, J.V. & Hurst, L.D., 2005. Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. *Genome biology*, 6(9), p.R75.

Chen, W.-H. et al., 2016. Integration of multi-omics data of a genome-reduced bacterium: Prevalence of post-transcriptional regulation and its correlation with protein abundances. *Nucleic acids research*, 44(3), pp.1192–1202.

Cock, P.J.A. et al., 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* , 25(11), pp.1422–1423.

Deutscher, M.P., 2006. Degradation of RNA in bacteria: comparison of mRNA and stable RNA. *Nucleic acids research*, 34(2), pp.659–666.

Eddy, S.R., 2011. Accelerated Profile HMM Searches. *PLoS computational biology*, 7(10), p.e1002195.

Fisher, R.A., 1922. On the Interpretation of χ2 from Contingency Tables, and the Calculation of P. *Journal of the Royal Statistical Society*, 85(1), pp.87–94.

Gardner, P.P. et al., 2011. Rfam: Wikipedia, clans and the "decimal" release. *Nucleic acids research*, 39(Database issue), pp.D141–5.

Giannoukos, G. et al., 2012. Efficient and robust RNA-seq process for cultured bacteria and complex community transcriptomes. *Genome biology*, 13(3), p.R23.

Goodman, D.B., Church, G.M. & Kosuri, S., 2013. Causes and effects of N-terminal codon bias in bacterial genes. *Science*, 342(6157), pp.475–479.

Hoeppner, M.P., Gardner, P.P. & Poole, A.M., 2012. Comparative Analysis of RNA Families Reveals Distinct Repertoires for Each Domain of Life C. O. Wilke, ed. *PLoS computational biology*, 8(11), p.e1002752.

Ikemura, T., 1981. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: A proposal for a synonymous codon choice that is optimal for the E. coli translational system. *Journal of molecular biology*, 151(3), pp.389–409.

Ioannidis, J.P.A. & Khoury, M.J., 2011. Improving validation practices in “omics” research. *Science*, 334(6060), pp.1230–1232.

Kudla, G. et al., 2009. Coding-sequence determinants of gene expression in Escherichia coli. *Science* , 324(5924), pp.255–258.

Kwon, T. et al., 2014. Protein-to-mRNA ratios are conserved between Pseudomonas aeruginosa strains. *Journal of proteome research*, 13(5), pp.2370–2380.

Laurent, J.M. et al., 2010. Protein abundances are more conserved than mRNA abundances across diverse taxa. *Proteomics*, 10(23), pp.4209–4212.

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

Lorenz, R. et al., 2011. ViennaRNA Package 2.0. *Algorithms for molecular biology: AMB*, 6, p.26.

Lu, P. et al., 2007. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nature biotechnology*, 25(1), pp.117–124.

Maier, T. et al., 2011. Quantification of mRNA and protein and integration with protein turnover in a bacterium. *Molecular systems biology*, 7, p.511.

Mückstein, U. et al., 2006. Thermodynamics of RNA-RNA binding. *Bioinformatics* , 22(10), pp.1177–1182.

Nakahigashi, K. et al., 2014. Effect of codon adaptation on codon-level and gene-level translation efficiency in vivo. *BMC genomics*, 15, p.1115.

Pain, A. et al., 2015. An assessment of bacterial small RNA target prediction programs. *RNA biology*, 12(5), pp.509–513.

Pelletier, J. & Sonenberg, N., 1987. The involvement of mRNA secondary structure in protein synthesis. *Biochemistry and cell biology = Biochimie et biologie cellulaire*, 65(6), pp.576–581.

Plotkin, J.B. & Kudla, G., 2011. Synonymous but not the same: the causes and consequences of codon bias. *Nature reviews. Genetics*, 12(1), pp.32–42.

Rozov, A. et al., 2015. Structural insights into the translational infidelity mechanism. *Nature communications*, 6, p.7251.

Schmittgen, T.D. & Livak, K.J., 2008. Analyzing real-time PCR data by the comparative C(T) method. *Nature protocols*, 3(6), pp.1101–1108.

Sharma, C.M. et al., 2010. The primary transcriptome of the major human pathogen Helicobacter

pylori. *Nature*, 464(7286), pp.250–255.

Sharp, P.M. & Li, W.-H., 1987. The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic acids research*, 15(3), pp.1281–1295.

de Sousa Abreu, R. et al., 2009. Global signatures of protein and mRNA expression levels. *Molecular bioSystems*, 5(12), pp.1512–1526.

Storz, G., Vogel, J. & Wassarman, K.M., 2011. Regulation by small RNAs in bacteria: expanding frontiers. *Molecular cell*, 43(6), pp.880–891.

Taniguchi, Y. et al., 2010. Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science*, 329(5991), pp.533–538.

Tuller, T. & Zur, H., 2015. Multiple roles of the coding sequence 5' end in gene expression regulation. *Nucleic acids research*, 43(1), pp.13–28.

Vivancos, A.P. et al., 2010. Strand-specific deep sequencing of the transcriptome. *Genome research*, 20(7), pp.989–999.

Vogel, C. & Marcotte, E.M., 2012. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature reviews. Genetics*, 13(4), pp.227–232.

Waters, L.S. & Storz, G., 2009. Regulatory RNAs in bacteria. *Cell*, 136(4), pp.615–628.

Workman, C. & Krogh, A., 1999. No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. *Nucleic acids research*, 27(24), pp.4816–4822.

Wu, D., Jospin, G. & Eisen, J.A., 2013. Systematic identification of gene families for use as markers for phylogenetic and phylogeny- driven ecological studies of bacteria and archaea and their major subgroups. *PLoS ONE*, 8(10), p.e77033.

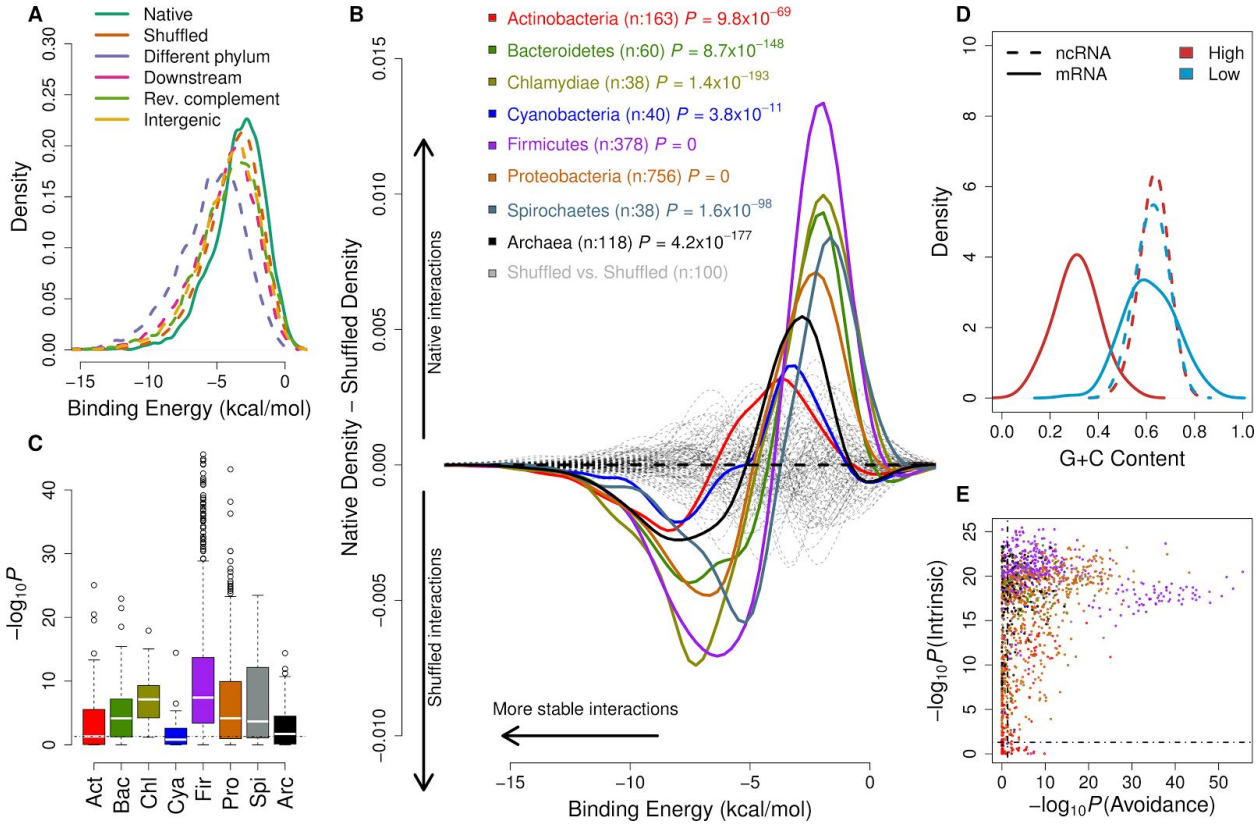

Figure 1. mRNA:ncRNA avoidance is a conserved feature of bacteria and archaea. **(A)** Native core mRNA:ncRNA binding energies (green line; mean = -3.21 kcal/mol) are significantly higher than all mRNA negative control binding energies (dashed lines; mean binding energies are -3.62, -5.21, -4.13, -3.86 & -3.92 kcal/mol respectively) in pairwise comparisons ($P < 2.2\text{x}10^{-16}$ for all pairs, one-tailed Mann–Whitney U test) for *Streptococcus suis* RNAs. **(B)** The difference between the density distributions of native mRNA:ncRNA binding energies and dinucleotide preserved shuffled mRNA:ncRNA controls as a function of binding energy for different taxonomic phyla. Each coloured curve illustrates the degree of extrinsic avoidance for different bacterial phyla or the archaea. Positive differences indicate an excess in native binding for that energy value, negative differences indicate an excess of interactions in the shuffled controls. The dashed black line shows the expected result if no difference exists between these distributions and the dashed grey lines show empirical differences for shuffled vs shuffled densities from 100 randomly selected bacterial strains. **(C)** This box and whisker plot shows $-\log_{10}(P)$ distributions for each phylum and the archaea, the P-values are derived from a

one-tailed Mann–Whitney U test for each genome of native mRNA:ncRNA versus shuffled mRNA:ncRNA binding energies. The black dashed line indicates the significance threshold ($P < 0.05$). **(D)** A high intrinsic avoidance strain (*Thermodesulfobacterium sp.* OPB45) shows a clear separation between the G+C distribution of mRNAs and ncRNAs ($P = 9.2\text{x}10^{-25}$, two-tailed Mann-Whitney U test), and a low intrinsic avoidance strain (*Mycobacterium sp.* JDM601) has no G+C difference between mRNAs and ncRNAs ($P = 0.54$, two-tailed Mann-Whitney U test). **(E)** The x-axis shows $-\log_{10}(P)$ for our test of extrinsic avoidance using binding energy estimates for both native and shuffled controls, while the y-axis shows $-\log_{10}(P)$ for our intrinsic test of avoidance based upon the difference in G+C contents of ncRNAs and mRNAs. Two perpendicular dashed black lines show the threshold of significance for both avoidance metrics. 97% of bacteria and archaea are significant for at least one of these tests of avoidance.

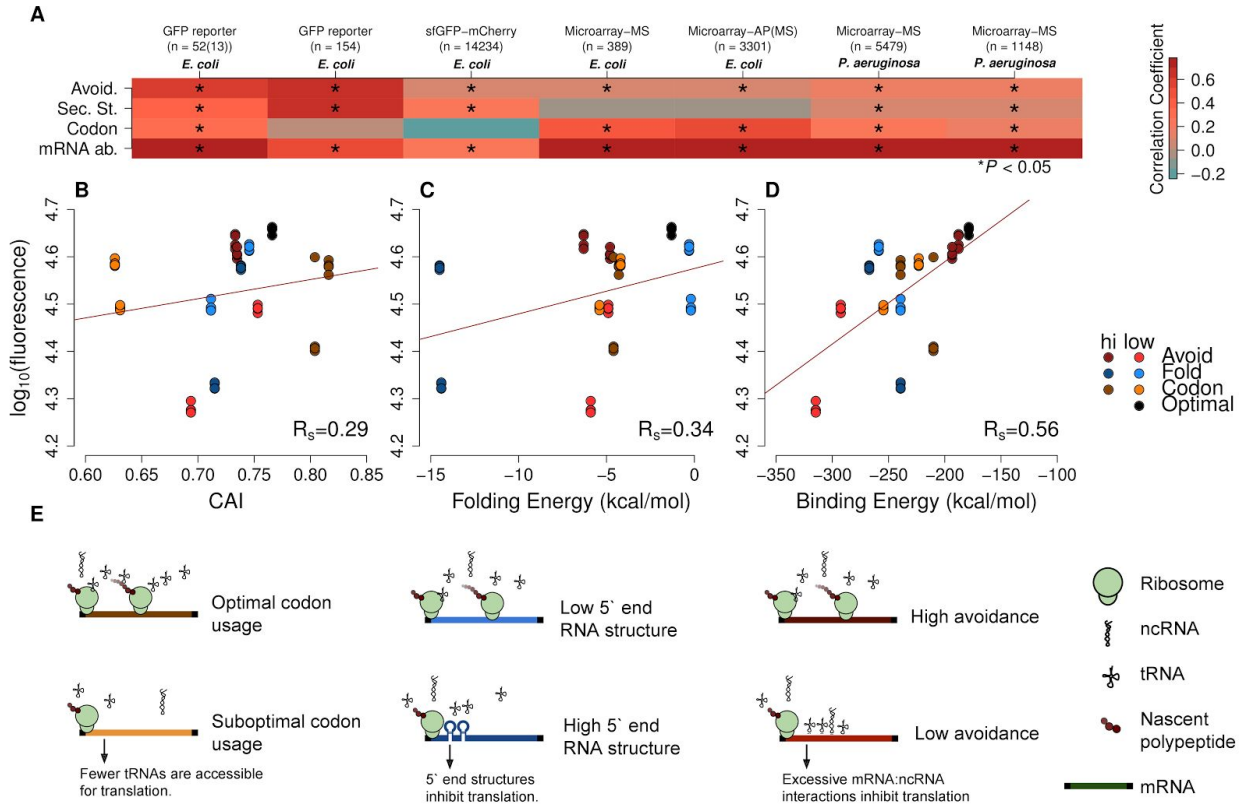

Figure 2. mRNA attributes have different impacts on protein abundance. **(A)** This heatmap summarizes the effect sizes of four mRNA attributes (avoidance of mRNA:ncRNA interaction, 5' end secondary structure, codon bias and mRNA abundance) on protein expression as Spearman's correlation coefficients, which are represented in gradient colors, while a starred block shows if the associated correlation is significant ($P < 0.05$). **(B)** GFP expression correlates with optimized codon selection, measured by CAI ($R_s = 0.29$, $P = 0.016$). **(C)** GFP expression correlates with 5` end secondary structure of mRNAs, measured by 5' end intramolecular folding energy ($R_s = 0.34$, $P = 0.006$). **(D)** GFP expression correlates with avoidance, measured by mRNA:ncRNA binding energy ($R_s = 0.56$, $P = 6.9\text{x}10^{-6}$). **(E)** Each cartoon illustrates the corresponding hypothesis; (1) optimal codon distribution (corresponding tRNAs are available for translation), (2) low 5` end RNA structure (high folding energy of 5` end) and (3) avoidance (fewer crosstalk interactions) lead to faster translation.

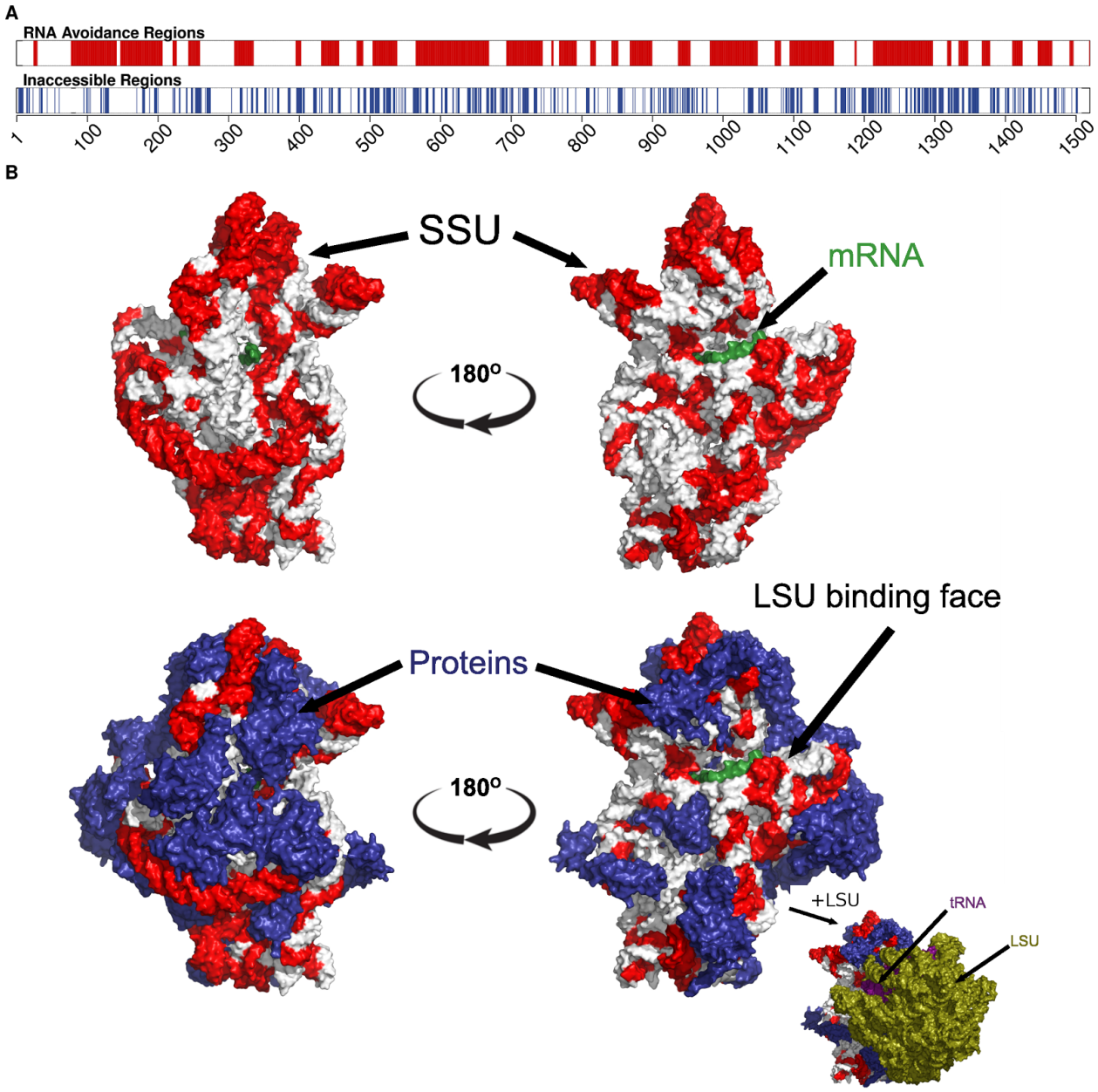

Figure 3. The most avoided regions of ribosomal SSU are related to nucleotide accessibility. **(A)** The red marked regions (top panel) correspond to *Thermus thermophilus* HB8 SSU rRNA regions that have a significant underrepresentation of stable mRNA:ncRNA interactions compared to randomized controls. The blue marked regions (bottom panel) correspond to inaccessible SSU rRNA nucleotides, which are in contact with other nucleotides or residues. **(B)** In order to identify the determinants of rRNA macromolecular structure that have a strong avoidance signal, we mapped the 2D data (A) onto the *T. thermophilus* SSU ribosomal RNA

structure (PDB ID: 4WZO) (Rozov et al. 2015). In this analysis, we have used a significance threshold of $P < 0.001$ to signify regions with avoidance (highlighted in red). When SSU ribosomal binding proteins and the large subunit (LSU) are to the view, these largely cover regions that do not have an avoidance signal (seen in white). We used a contingency table to test whether avoidance and binding to other molecules (i.e. LSU rRNA and ribosomal proteins) were related. We found a statistically significant relationship between low avoidance regions and those that are protected by other molecules on SSU rRNA ($P=2.5x10^{-17}$, Fisher's exact test). This indicates that the most accessible regions of ribosomal RNA are those that have the largest impact on mRNA sequence-space.

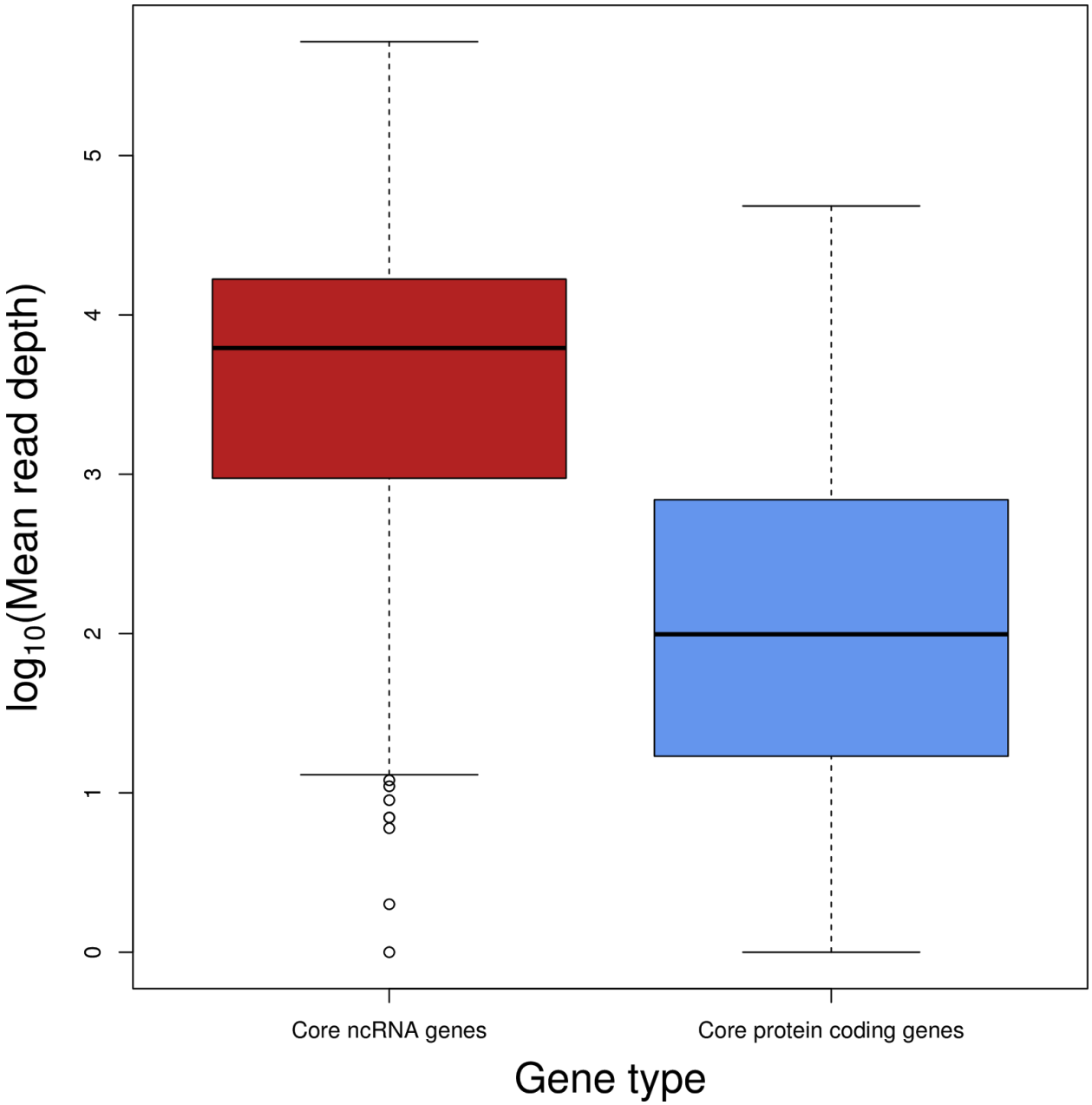

Figure 4. The median expression of core ncRNA genes (n=325 data points) in prokaryotic genomes is nearly two orders of magnitude greater than core mRNAs (n=8086 data points) which proves that ncRNAs constitute most of the cellular RNAs. To create this plot, we used mean mapped reads per gene length (i.e. mean read depth per position) of each core gene. The expression data is compiled from 5 archaeal and 37 bacterial strains from a previous study (Lindgreen et al. 2014).

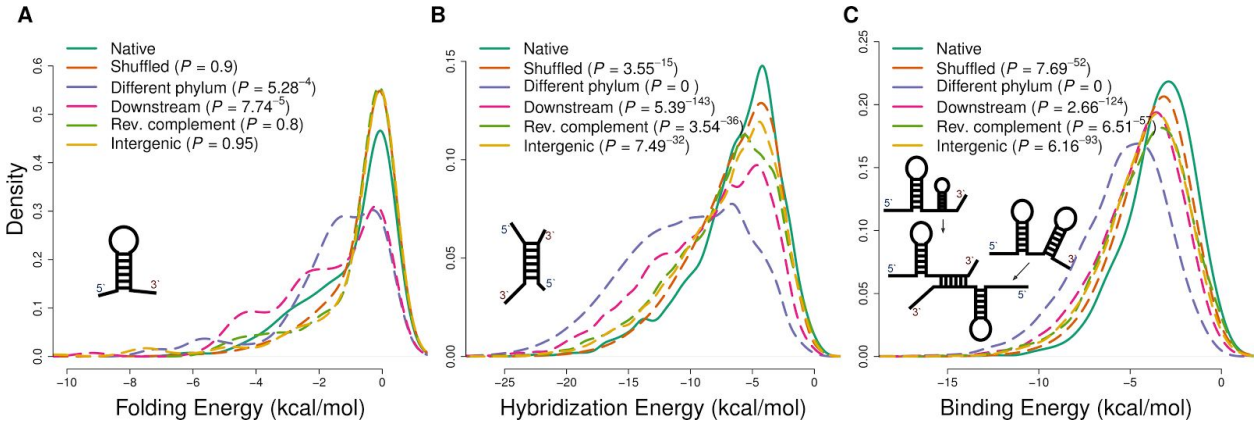

Figure 1-figure supplement 1. Applying different energy models of intramolecular and intermolecular interactions for native sequences and various negative controls. **(A)** The distributions of internal secondary structure (intramolecular) minimum free energies (MFEs) for 5` ends of mRNA sequences, estimated using RNAfold from the Vienna package (Lorenz et al. 2011). **(B)** The distributions of hybridisation MFEs between core mRNAs and ncRNAs, estimated using the RNAduplex algorithm from the Vienna package (Lorenz et al. 2011). **(C)** The distributions of binding MFEs between core mRNAs and ncRNAs, estimated using the RNAup algorithm (Lorenz et al. 2011). The RNAup algorithm minimizes the sum of energies necessary to open binding sites on two RNA molecules and the hybridization energy (Lorenz et al. 2011). This method has been shown to be the most accurate general approach for sequence-based RNA interaction prediction (Pain et al. 2015).

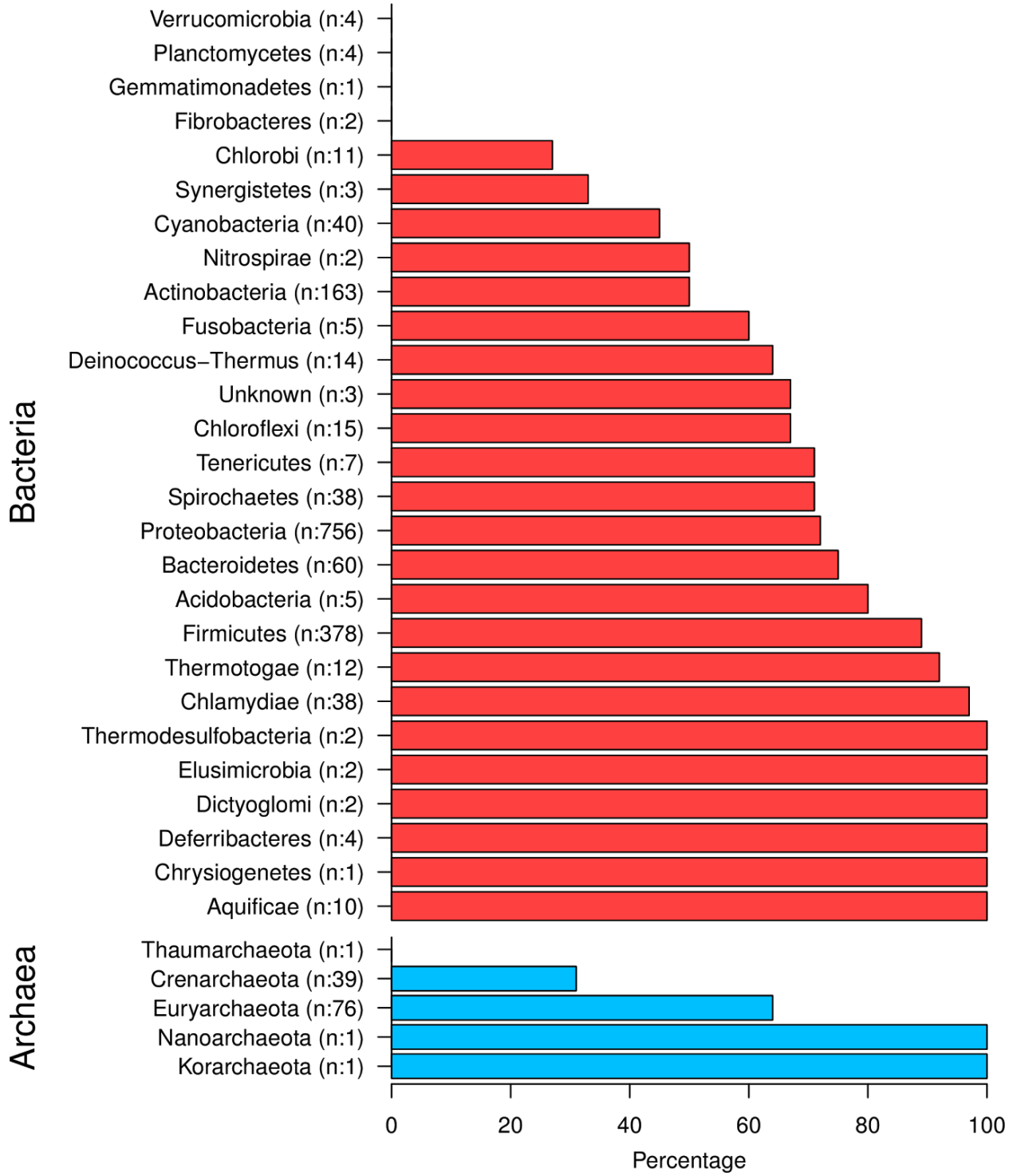

Figure 1-figure supplement 2. The top and the bottom panels shows bacterial phyla and archaeal phyla respectively. Numbers in brackets show the total members and the x-axis displays the percentage of extrinsic avoidance conservation in associated phylum. The archaeal and bacterial phyla with fewer than 20 publicly available sequenced genomes were excluded from

further analysis due to concerns about sample size sufficiency.

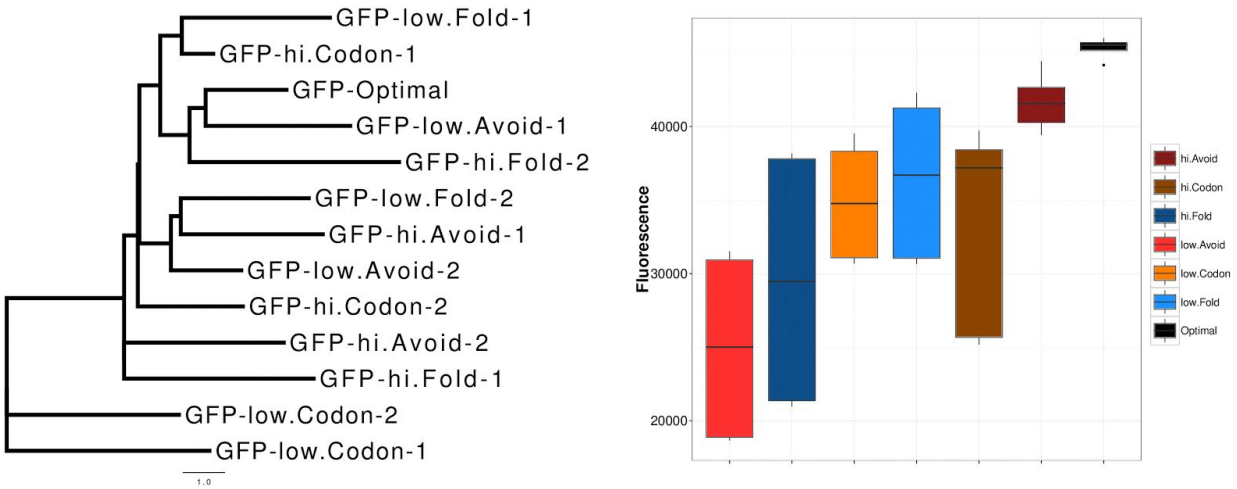

Figure 2-figure supplement 1. GFP mRNA constructs have unbiased design that produce different protein expressions. An unrooted maximum likelihood tree of the extreme GFP mRNAs on the left panel illustrates the low similarity between our GFP mRNA constructs. The distances were calculated using HKY85 nucleotide substitution model. On the right panel, the y-axis shows relative fluorescence units (RFU) of GFP expression from synonymously sampled mRNAs with different characteristics, these are labelled on the figure legend. Optimal and high avoidance GFP mRNAs produce the highest expression while low avoidance GFP mRNAs have the lowest expression ($P = 1.35\text{x}10^{-5}$, Kruskal-Wallis test).

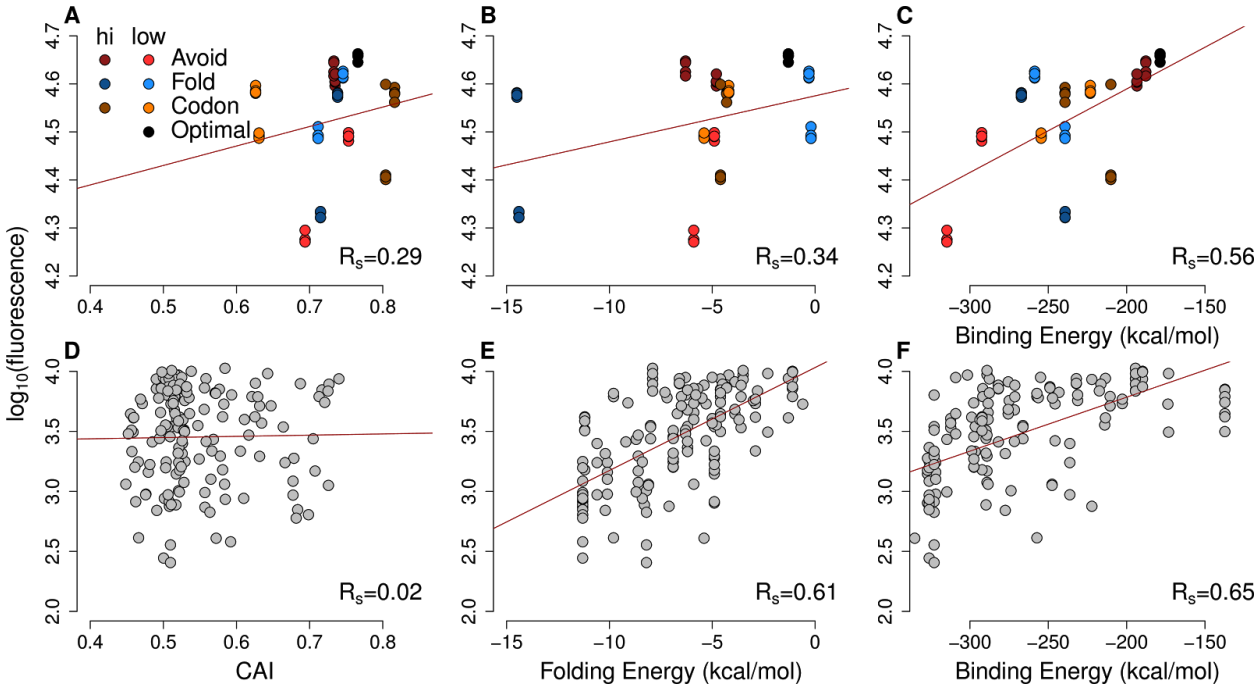

Figure 2-figure supplement 2. The scatter-plots of protein abundances (as log-fluorescences) summarize the effect of general factors for extreme GFP and previously published GFP datasets. **(A) (B) (C)** Each GFP mRNA was sampled from the extremes of one of three metrics presumed to impact expression mRNA:ncRNA binding, 5` end secondary structure or codon usage. Slightly darker or lighter colors display the type of extremes. Avoidance correlates with GFP expression ($R_s$ = 0.56, $P$ = 6.9x10$^{-6}$) more than CAI ($R_s$ = 0.29, $P$ = 0.01) and 5` end folding energy (R_s = 0.34, P = 0.006). **(D) (E) (F)** Using a previously published GFP dataset (Kudla et al. 2009) the CAI does not correlate with protein abundance (R_s = 0.02, P = 0.4), while 5` end folding energy (R_s = 0.61, P = 5.7x10^{-18}) and avoidance (R_s = 0.65, P = 1.6x10^{-20}) influence GFP expression.

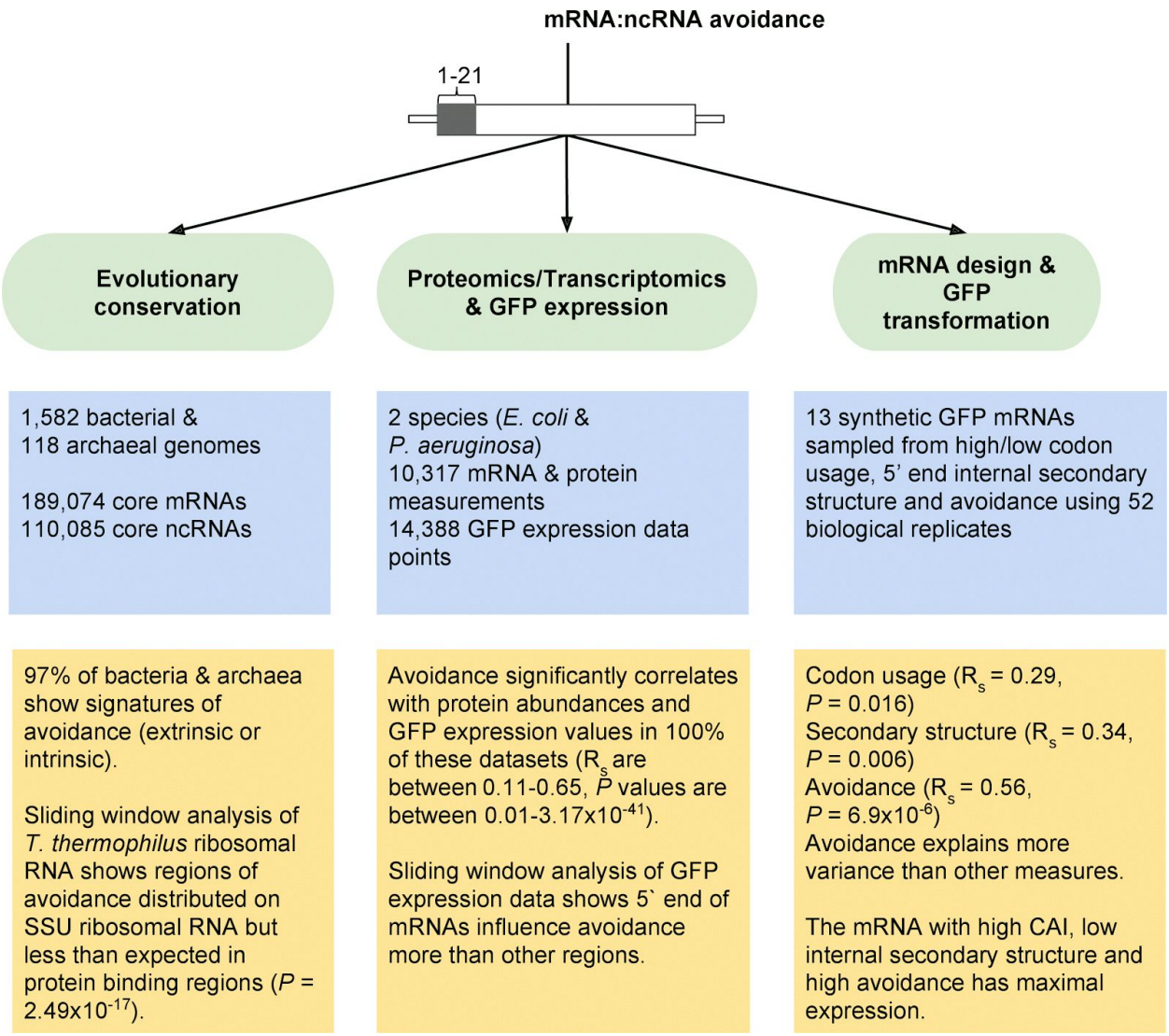

Figure 2-figure supplement 3. Overview of mRNA:ncRNA avoidance analysis and results. Our tests for avoidance can be divided into three main parts; (1) evolutionary conservation analyses to detect energy shifts in bacterial and archaeal genomes relative to dinucleotide shuffled negative controls, (2) analyses of proteomics, transcriptomics and GFP transformation data to predict the effect size of avoidance on protein expression and lastly (3) the application of avoidance hypothesis to design synonymous mRNAs that either produce high or low levels of corresponding protein.

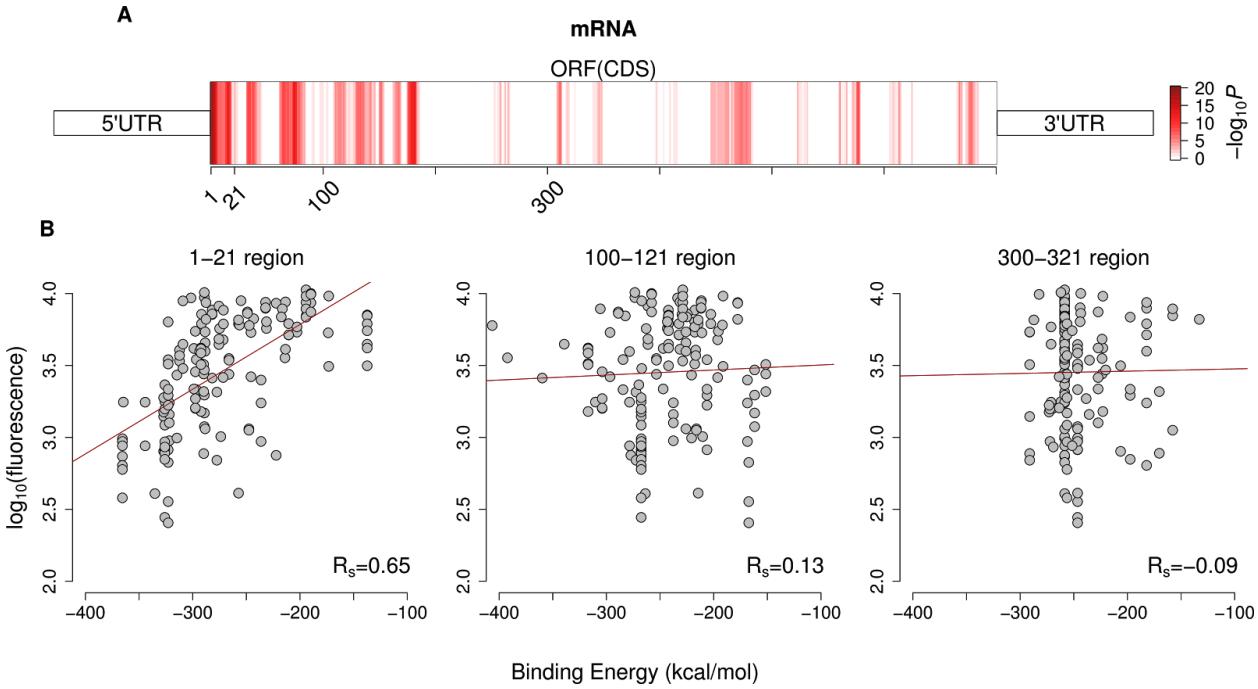

Figure 3-figure supplement 1. Avoidance pattern and its correlation with protein expression vary on mRNAs. **(A)** A sliding window (length 21, step size 1) analysis based on previously published GFP expression dataset (Kudla et al. 2009) shows the significance of correlation between avoidance and their corresponding fluorescence values for each position along the coding region. Darker red regions show more significant positions (with higher -$\log_{10}P$ values). **(B)** This analysis proves that binding energy of first 21 nt region influences protein expression more than any other downstream region and corresponding Spearman's correlation coefficients for selected sliding window start positions are seen at bottomright. It also justifies our selection of 5` end coding region for avoidance.

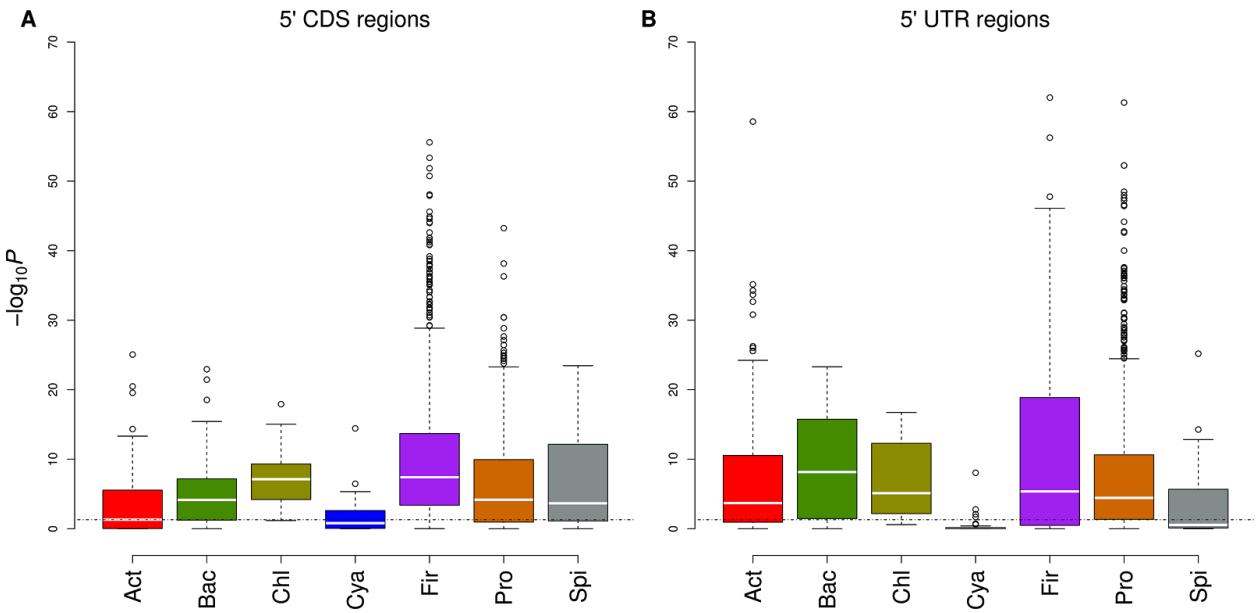

Figure 3-figure supplement 2. Comparison of different regions for evolutionary conservation analyses. **(A)** This box and whisker plot (similar with Figure 1C except archaea) shows -$\log_{10}(P)$ distributions for each bacterial phylum. The black dashed line indicates the significance threshold ($P < 0.05$). We used 5` end CDS regions as designated interaction location. **(B)** In this plot, 5` end UTR regions (90 nucleotides upstream to 21 nucleotides downstream) are used as designated interaction regions. It seems both regions have similar avoidance conservation, which proves avoidance is not limited to 5` ends of coding region.