

RNA sectors and allosteric function within the ribosome

Allison S. Walker^{1,*}, William P. Russ³, Rama Ranganathan^{4,*}, and Alanna Schepartz^{1,2,*}

¹ Department of Chemistry, Yale University, New Haven, CT 06520

² Department of Molecular, Cellular, and Developmental Biology, Yale University, New Haven, CT 06520

³ Green Center for Systems Biology, UT Southwestern Medical Center, Dallas, TX 75390

⁴ Center for Physics of Evolving Systems, Biochemistry & Molecular Biology, and the Institute for Molecular Engineering, The University of Chicago, Chicago, IL 60637

* To whom correspondence should be addressed. Tel: 1-203-432-5094; Fax: 203-432-3486; Email: schepartz@berkeley.edu; Correspondence may also be addressed to allison_walker@hms.harvard.edu, ranganathanr@uchicago.edu

Supplementary methods

Assembly of multiple sequence alignment (MSA).

We used a custom python script to generate an MSA using sequences from the comparative RNA web (CRW) (1). The CRW has separate alignments for bacteria, archaea, eukaryotes, mitochondria, and chloroplasts (2,944 sequences in total). Each alignment contains a reference *E. coli* sequence to which all sequences are aligned. Our script used the *E. coli* reference sequence to create a complete MSA in which all sequences are aligned to that of *E. coli*. To shorten the length of the alignment, our script removed from the alignment all positions corresponding to a gap in the *E. coli* sequence. We used this MSA as the input for both SCA and DCA.

Calculation of the SCA coupling matrix

Our modification of the SCA algorithm to evaluate RNA sequences in place of protein began with a previously described Python implementation of SCA (SCA version 6). We modified the SCA 6 code to use background frequencies of nucleic acids rather than amino acids, which we calculated from our MSA. These frequencies were: A = 0.2706, G = 0.306, U = 0.207, C = 0.217.

Prior to calculating the SCA coupling matrix as described below, we filtered the MSA to remove any position that corresponded to a gap in at least 80% of all sequences. We then filtered the MSA to remove any sequence in which 20% or more of the positions in the alignment corresponded to a gap or any sequence that was less than or equal to 20% identical to the *E. coli* sequence. We then filtered the positions a second time, removing any position that corresponded to a gap in at least 20% of all sequences. Finally, we weighted each sequence by a factor corresponding to the inverse of the number of sequences in the alignment with which it shared more than 80% sequence identity. We first calculated the unweighted coupling using Equation (1),

$$C_{ij}^{ab} = f_{ij}^{ab} - f_i^a f_j^b \quad \text{Equation (1)}$$

where f_{ij}^{ab} is the joint frequency of nucleotides *a* and *b* at positions *i* and *j* and f_i^a is the frequency of nucleotide *a* at position *i*, such that the conservation weighted statistical coupling is defined by Equation (2).

$$\tilde{C}_{ij}^{ab} = \phi_i^a \phi_j^b C_{ij}^{ab} \quad \text{Equation (2)}$$

Here, ϕ_i^a represents the conservation at position *i* as calculated by the first derivative of the Kulback-Leibler entropy (Equation 3):

$$D_i^a = f_i^a \ln \frac{f_i^a}{q^a} + (1 - f_i^a) \ln \frac{1 - f_i^a}{1 - q^a}$$
$$\phi_i^a = \left| \frac{\partial D_i^a}{\partial f_i^a} \right| \quad \text{Equation (3)}$$

Here, f_i^a represents the frequency of nucleotide a at position i and q^a represents the background frequency of the nucleotide. We then reduced the $L \times L \times 4 \times 4$ matrix to an $L \times L$ matrix using the Frobenius norm (2) of each 4×4 matrix.

SCA IC determination.

We determined the SCA Independent Components (ICs) as previously described (3). We used spectral decomposition (2) to calculate the eigenvalues and eigenvectors of the $L \times L$ coupling matrix. We determined the statistical significance of eigenvectors by comparing the calculated eigenvalues to the eigenvalues of the statistical coupling matrices of ten randomly generated MSAs with the same number of sequences and same frequencies of nucleotides as the 23S rRNA MSA. An eigenvector was considered statistically significant if the eigenvalue was greater than the mean of the 1st random eigenvalues plus three times the standard deviation of the random eigenvalues. Then we applied independent component analysis (4) to the statistically significant eigenvectors to maximize their independence and calculate the ICs. In order to define which bases were in each IC, we fit the entries of each independent component (IC) to a t-distribution. We considered positions with entries in the top 95th percentile of the t-distribution to be part of the IC. If a position was found in the top 95th percentile or above for multiple ICs, then we assigned it to the IC to which it was most strongly coupled.

SCA sector determination.

To define sectors, we examined the extent of inter-IC coupling. We calculated the inter-IC coupling as the average coupling using Equation (4),

$$\text{coupling}(IC_i, IC_j) = \frac{1}{n_i n_j} \sum_{k \in \text{residues in } IC_i} \sum_{l \in \text{residues in } IC_j} \tilde{C}_{kl} \quad \text{Equation (4)}$$

where n_i is the number of bases in IC_i , n_j is the number of bases in IC_j , and \tilde{C} is the $L \times L$ coupling matrix. We considered any pair of ICs with an inter-IC coupling greater than or equal to the average plus two standard deviations of all inter-IC couplings to be part of the same sector.

Generation of 23S mutant library

Each Gibson Assembly reaction required two PCR products: a segment of the PLK35 plasmid backbone lacking the region surrounding the desired mutation and an insert containing the mutated sequence. All primer sequences are listed in the Table S5. PCR was performed using Phusion™ DNA polymerase (NEB). Each reaction was optimized with respect to annealing temperature (55-70°C), template DNA concentrations (10-100 ng/50 µL reaction), and number of cycles (25-30 cycles). The products were digested with DpnI (New England Biolabs) and the DNA purified (Qiagen, PCR purification kit). Gibson assembly reactions (5) were performed using standard protocols (Gibson Assembly Kit, New England Biolabs). We individually transformed plasmids containing the desired mutation into SQZ10 cells (6) and plated the cells on agar containing 100 µg/mL carbenicillin and 8% sucrose. We generated three glycerol stocks of each mutant which were stored at -80°C until needed for continuous culture experiments. We also generated glycerol stocks of the SQZ10 strain transformed with a PLK35 plasmid harboring a WT 23S rRNA gene.

All mutations were confirmed by Sanger sequencing (Quintara Biosciences). From Sanger sequencing, we found that some of the Gibson Assemblies resulted in mutations near the intended location. We observed mutations at 2079 produced by Gibson Assemblies designed to install mutations in the 2056-2064 region and mutations at 2110, 2154, and 2160 produced by Gibson Assemblies designed to install mutations in the 2112-2115 region. We observed mutations at 2465 produced by Gibson Assemblies designed to install mutations in the 2483-2507 region. We included these mutants in our library.

Continuous culture experiments.

Approximately 1 µL of each glycerol stock was added to each well of a 96-well plate containing 200 µL of Luria-Bertani Broth (LB) supplemented with 100 µg/mL carbenicillin (LBC) and the plates were incubated overnight at 37°C. The resulting cultures were combined and diluted with LBC to an OD₆₀₀ of 0.1. We then added 150 mL of this diluted culture to a turbidostat maintained at room temperature and set to clamp the OD₆₀₀ of the culture to 0.1. Whenever the culture reached an OD₆₀₀ higher than 0.1, the turbidostat diluted the culture by adding LBC. We removed 5 mL samples of the continuous culture from the turbidostat after 0, 18, 20, 24, and 26 h; the samples were then

centrifuged and plasmid DNA extracted (Qiagen Miniprep Kit). We performed two turbidostat runs simultaneously and in parallel starting from the same set of 96 well plates. We repeated the parallel turbidostat runs to generate a total of four data sets for analysis. We used PCR to amplify and barcode the region of interest for illumine sequencing. A full description of PCR protocols is available in the Supplementary Information.

Next-generation sequencing.

In order to sequence the samples collected during the continuous culture experiments, we first PCR amplified the 2039-2550 region of the 23S operon from the plasmid DNA obtained at each time point using forward and reverse PCR primers (listed in the supplementary data) purchased from IDT and the KOD DNA polymerase (Millipore). The sequence was amplified over 30 PCR cycles. Each primer contained a stretch of six Ns to reduce redundancy at the beginning of a sequencing read and improve read quality (7). Each primer also contained a region complementary to the Illumina adapter primers (underlined in the sequences above) to enable a second round of PCR to append the Illumina Adapter Sequences.

We used the products from the first PCR reaction as templates for a second round of PCR to append the Illumina TruSeq Adapter Sequences. For each sequencing run, we PCR amplified each sample with primers containing a unique combination of forward and reverse barcodes (purchased from IDT, primers listed in the Table S4). We then gel purified (Zymogen Gel DNA recovery kit) each amplicon and sequenced them. MiSeq sequencing was performed by Quintara Biosciences and HiSeq sequencing was performed by the Yale Center for Genome Analysis (YCGA). Because the samples being sequenced had very low sequence diversity, Quintara added PhiX (Illumina, for MiSeq runs) and the YCGA added human exome sequences (provided by YCGA) to improve read quality.

Primers used to amplify the 2039-2550 region of the 23S operon from the plasmid DNA for next generation sequencing:

Forward primer: TGACTGGAGTTCAGACGTGgtctctccgatctNNNNNNCAGTGTACCCGCGGCAAGAC

Reverse primer: CACTCTTTCCTACACGACGctctctccgatctNNNNNNCCATACCCTTGGGACCTACTTCAGC

Processing of next-generation sequencing data

Using our custom scripts, we first filtered out sequencing reads that lacked sequence corresponding to binding sites for the forward and reverse primers used to amplify the PLK35 plasmid 2039-2550 region: GCAAGAC for forward reads and AAGGGTAT for reverse reads. If either of these sequences were absent, the sequence was discarded. Due to the lower quality of HiSeq reads relative to MiSeq reads, we also discarded any HiSeq reads with an average quality score less than or equal to 20. We compared the first 230 bases of the reverse read and the first 240 bases of the forward read of all remaining sequences to the expected WT 23S rRNA sequence and stored the location of any apparent mutation(s) and the quality of the read at that position. To reduce the number of sequences processed in subsequent analyses, we discarded any sequence with more than 30 apparent mutations in either the forward or the reverse read. We then combined the forward and reverse read mutation lists. During this step we also discarded mutations that seemed to be the result of a sequencing error rather than a true mutation. To make this assessment, we first counted all instances of a mutation at a position. If the mutation occurred two times or fewer and the quality score at that position was less than or equal to 20, we considered the mutation to be a sequencing error rather than a true mutation. Additionally, because the HiSeq sequencing reads were of especially poor quality near the ends of the reads (Figure S4B), we discarded any mutations that occurred after position 2256 in the forward read or before 2400 in the reverse read. We consider it unlikely that this process inadvertently discarded any true mutations because we did not introduce any mutants in this region and did not observe any unanticipated mutations in this region when the DNA was sequenced by MiSeq. Finally, we counted all sequences that had zero mutations (WT) and all sequences with just a single mutation.

For each experiment, we discarded data for any mutants that were present in fewer than 300 reads at any of the time points. Then we used the WT and single mutant counts from each sample to calculate the relative frequency of each mutant using Equation 5:

$$f = \log\left(\frac{M_t}{M_0}\right) - \log\left(\frac{W_t}{W_0}\right) \quad \text{Equation (5)}$$

where M_t is the number of sequencing reads for mutant M at time t, M_0 is the number of sequencing reads for mutant M at time = 0, W_t is the number of WT sequencing reads at time t, and W_0 is the number of sequencing reads for the WT sequence at time = 0. We fit the frequency from each time point (excluding the zero time-point) with a linear regression and discarded any mutant where the regression was poor ($R^2 < 0.6$). We repeated this procedure for each sequencing runs of each experiment. Ultimately, we only included mutants that passed thresholds in at least two sequencing experiments.

Calculation of correlation between experiment and coupling predictions.

Graphs shown in Figure 4 and S5 were generated by plotting the relative growth rate of each mutant as a function of the weighted average (w_i defined in the formula below) of the contribution of the mutated residue to all ICs in the sector (the value associated with the mutated residue's position in the IC). The w_i values were calculated using the Equation 6:

$$w_i = IC_i^T C IC_i \quad \text{Equation (6)}$$

In this equation, IC_i is the vector corresponding to the i^{th} independent component and C is the LxL statistical coupling matrix. The weight (w_i) corresponds to the variance explained by each independent component.

We calculated the weighted average of the IC vectors using Equation 7:

$$\overline{IC} = \frac{1}{\sum_{i=1}^k w_i} \sum_{i=1}^k w_i IC_i \quad \text{Equation (7)}$$

In this equation, \overline{IC} is the weighted average of the IC vectors and k is the number of ICs in the sector. We calculated Pearson's correlation coefficients, R^2 values, and p values using PRISM (GraphPad).

Projection of ICs onto sequence space

The IC vectors for each sector can be projected onto sequence space as described in (8). First the statistically significant eigenvectors are projected onto sequence space using the formula: $U = XVA^{-1/2}$ where U is a matrix containing the projected eigenvectors, X is an M x L matrix representing the alignment, V is a matrix containing the 26 statistically significant sectors, and A is a diagonal matrix containing the 26 statistically significant eigenvalues (this formula follows from singular value decomposition). To obtain the projected IC vectors, the unmixing matrix derived during ICA is applied to U .

References cited

1. Cannone, J.J., Subramanian, S., Schnare, M.N., Collett, J.R., D'Souza, L.M., Du, Y., Feng, B., Lin, N., Madabusi, L.V., Muller, K.M. *et al.* (2002) The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics*, **3**, 2.
2. Golub, G.H. and Van Loan, C.F. (1996) *Matrix Computations*, 3rd ed. Johns Hopkins, Baltimore, MD.
3. Reynolds, K.A., Russ, W.P., Socolich, M. and Ranganathan, R. (2013) Evolution-Based Design of Proteins. *Methods Enzymol*, **523**, 213-235.
4. Bell, A.J. and Sejnowski, T.J. (1995) An information-maximization approach to blind separation and blind deconvolution. *Neural Comput*, **7**, 1129-1159.
5. Gibson, D.G., Young, L., Chuang, R.Y., Venter, J.C., Hutchison, C.A., 3rd and Smith, H.O. (2009) Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Methods*, **6**, 343-345.
6. Asai, T., Zaporozhets, D., Squires, C. and Squires, C.L. (1999) An Escherichia coli strain with all chromosomal rRNA operons inactivated: Complete exchange of rRNA genes between bacteria. *Proc Natl Acad Sci U S A*, **96**, 1971-1976.

7. Kou, R., Lam, H., Duan, H., Ye, L., Jongkam, N., Chen, W., Zhang, S. and Li, S. (2016) Benefits and Challenges with Applying Unique Molecular Identifiers in Next Generation Sequencing to Detect Low Frequency Mutations. *PLoS One*, **11**, e0146638.
8. Rivoire, O., Reynolds, K.A. and Ranganathan, R. (2016) Evolution-Based Functional Decomposition of Proteins. *PLoS Comput Biol*, **12**, e1004817.
9. Bernier, C.R., Petrov, A.S., Waterbury, C.C., Jett, J., Li, F., Freil, L.E., Xiong, X., Wang, L., Migliozi, B.L., HersHKovits, E. *et al.* (2014) RiboVision suite for visualization and analysis of ribosomes. *Faraday Discuss*, **169**, 195-207.
10. Petrov, A.S., Bernier, C.R., Hsiao, C.L., Norris, A.M., Kovacs, N.A., Waterbury, C.C., Stepanov, V.G., Harvey, S.C., Fox, G.E., Wartell, R.M. *et al.* (2014) Evolution of the ribosome at atomic resolution. *P Natl Acad Sci USA*, **111**, 10251-10256.
11. Petrov, A.S., Gulen, B., Norris, A.M., Kovacs, N.A., Bernier, C.R., Lanier, K.A., Fox, G.E., Harvey, S.C., Wartell, R.M., Hud, N.V. *et al.* (2015) History of the ribosome and the origin of translation. *P Natl Acad Sci USA*, **112**, 15396-15401.
12. Bokov, K. and Steinberg, S.V. (2009) A hierarchical model for evolution of 23S ribosomal RNA. *Nature*, **457**, 977-980.

Supplemental Figures and legends.

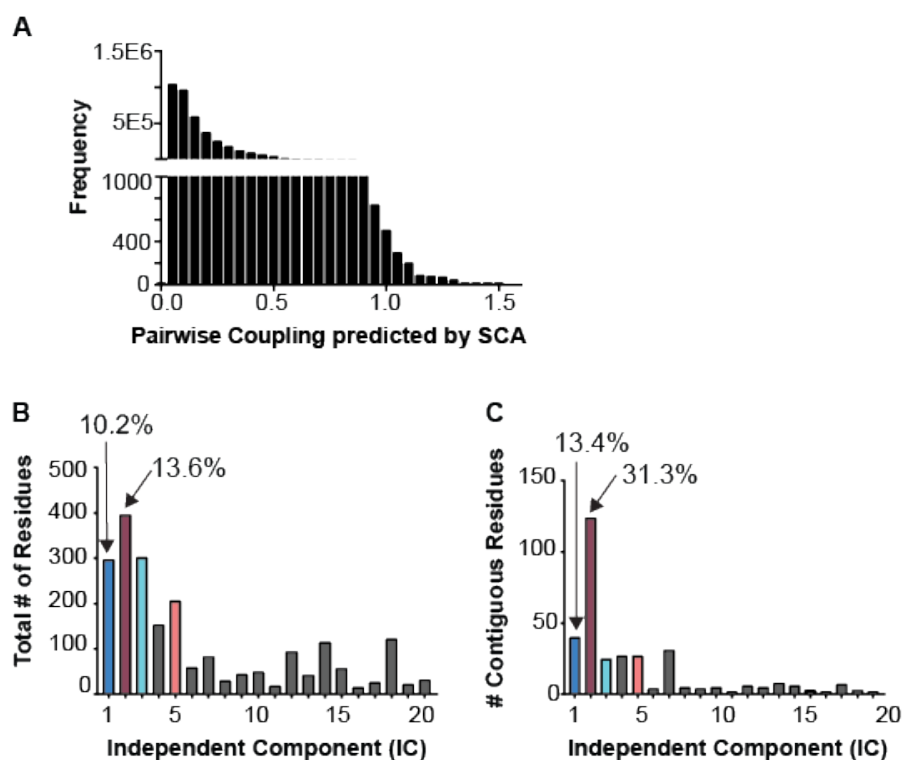


Figure S1. Pairwise coupling and ICs. Histogram showing the frequency distribution of coupling values calculated using SCA. The distribution reveals very high coupling values for only a small percentage of all nucleotide pairs in the alignment. (B) Bar graph showing the total number of residues within each of the top 20 ICs predicted by SCA. (C) Bar graphs showing the largest number of contiguous residues in each of the top 20 ICs predicted by SCA. A residue is defined as part of a contiguous group if any atom within that residue is located within 8 Å of any atom within at least one residue in the group. The percent values shown represent the fraction of positions in the IC that are contiguous.

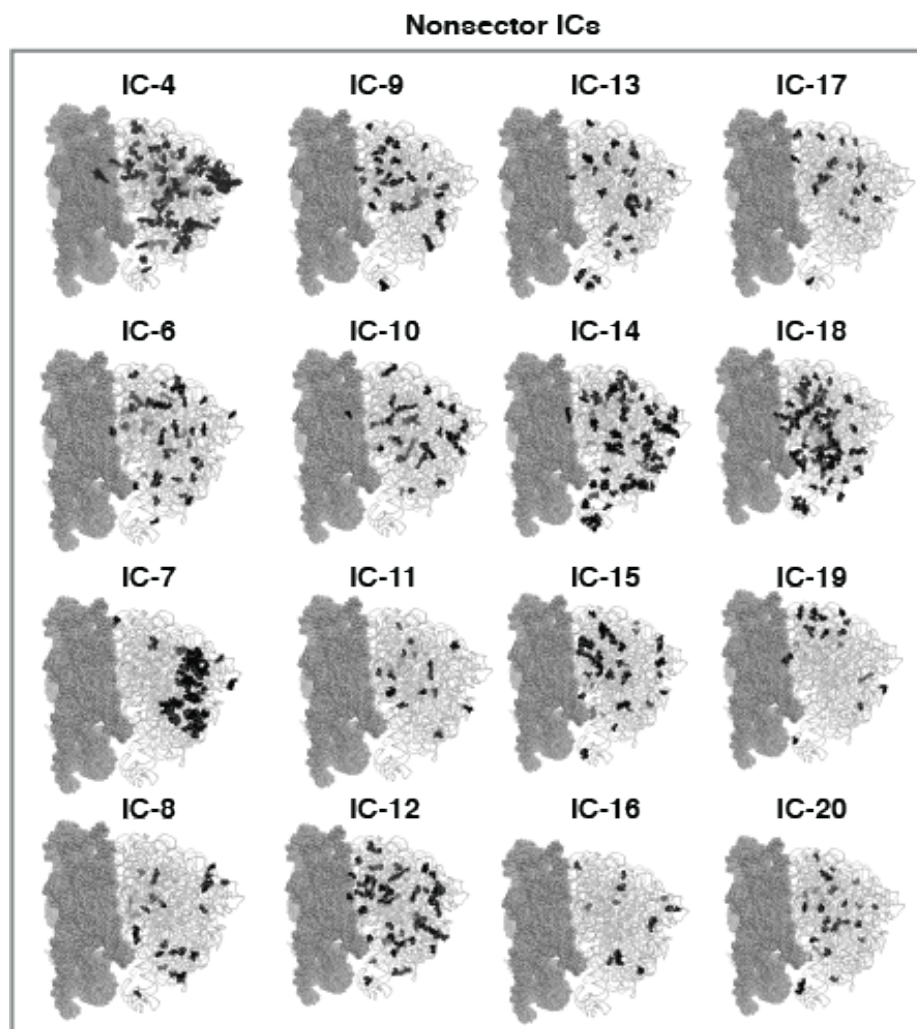


Figure S2. Spatial distribution of SCA-predicted ICs that are not depicted in Figure 1. Spatial distribution of statistically significant SCA-predicted ICs that are not part of a sector. The ICs are mapped onto the structure of the *E. coli* ribosome (PDB 5JTE).

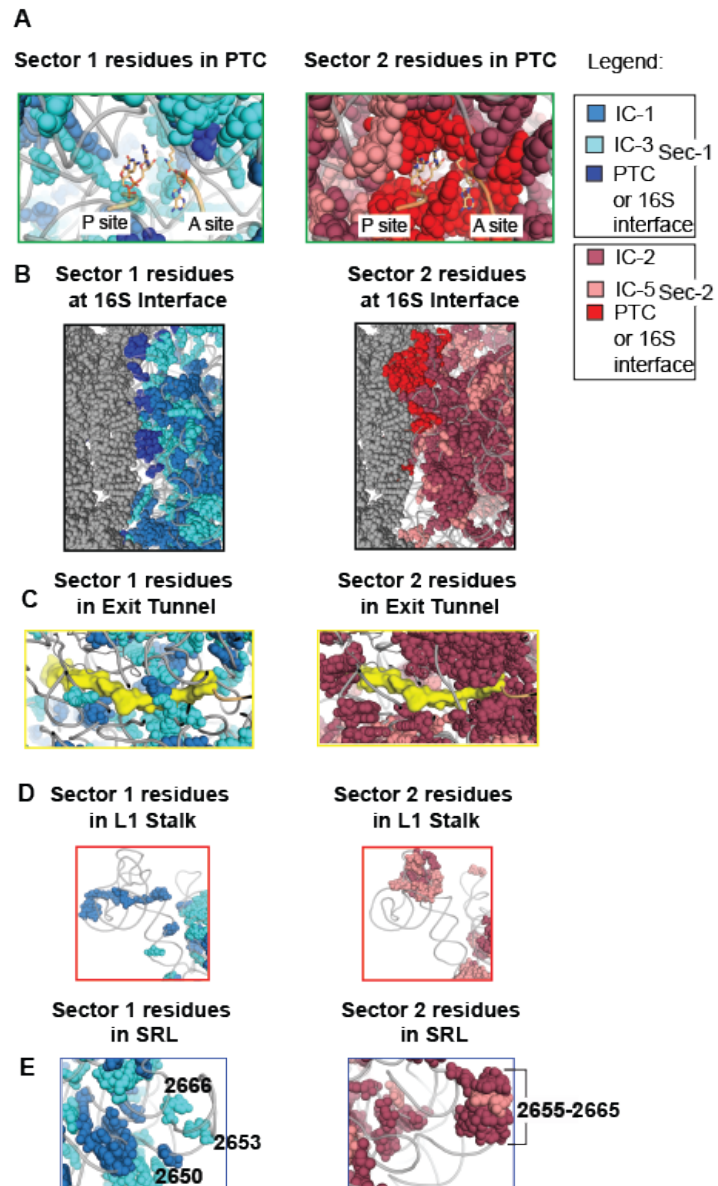


Figure S3. Sectors contain residues within multiple essential ribosome regions. Close-up view of the (A) PTC; (B) inter-subunit interface; (C) exit tunnel with the SecM peptide from the structure 3JBU shown as a yellow surface; (D) L1 stalk; and (E) SRL showing the locations of residues in Sector 1 (blues) and Sector 2 (pinks) (purples). In (A) and (B) residues that are in the sector and in the PTC or at the interface are shown in a bright blue (Sector 1) or bright red (Sector 2).

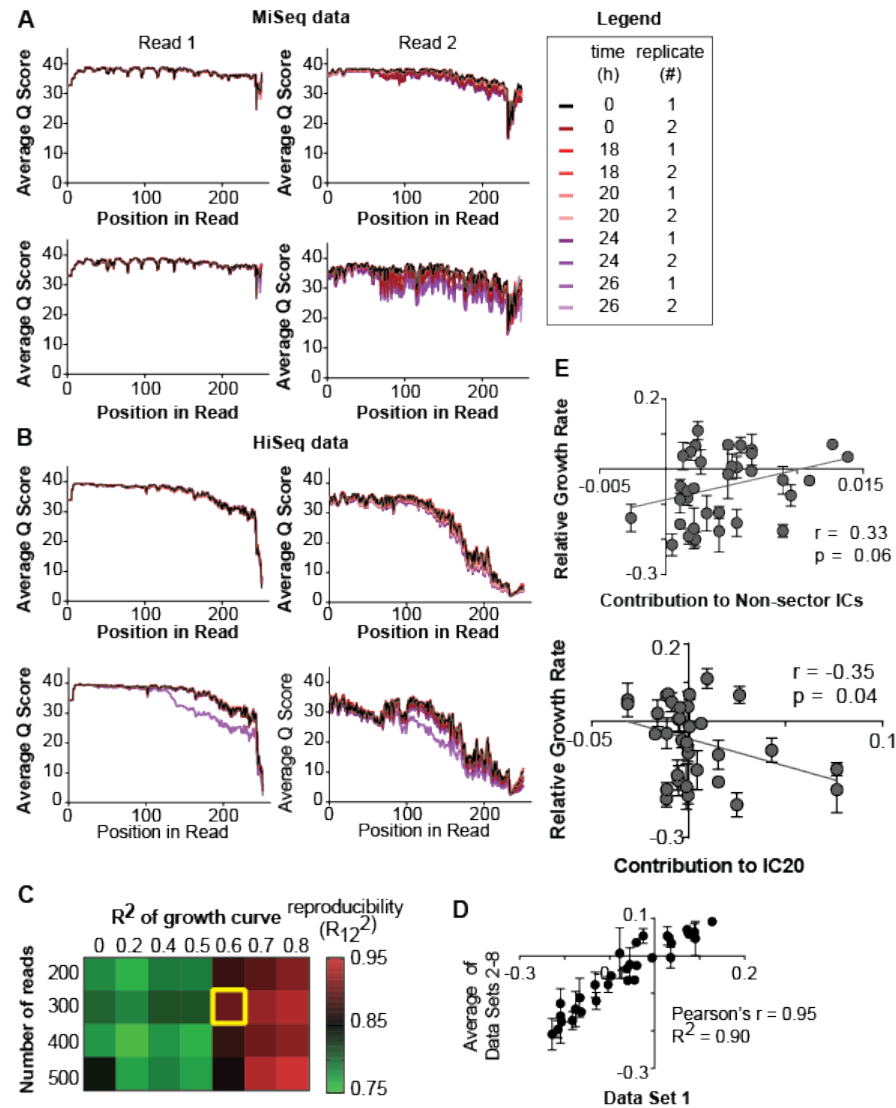


Figure S4. Processing and analysis of next-generation sequencing data. Plots show the average Illumina quality score at each position in the read during MiSeq (A) and HiSeq (B) sequencing of clones obtained at the indicated timepoint. Data for both read 1 (reverse read) and read 2 (forward read) are shown. (C) Heat map showing the relationship between the reproducibility of the experiment and two thresholds used for sequencing analysis: the minimum number of mutant reads and the minimum R^2 value of the growth curve (calculated from the linear regression of time vs log of enrichment relative to WT). Coloring indicates the reproducibility of the continuous culture experiment as measured by the correlation between relative growth rates calculated from two different continuous culture data sets (R_{12}^2). A redder coloring indicates a higher R_{12}^2 and therefore higher reproducibility between experiments. The yellow box indicates the set of thresholds we chose to use in our analysis. (D) Plot showing the reproducibility of the experiment using the chosen thresholds (minimum number of sequence reads for mutant = 300 and R^2 of growth curve = 0.6). The x-axis represents the growth rates of each mutant in one of our sequencing data sets and the y-axis represents the average growth rates of each mutant in the other seven sequencing data sets. (E) Plot showing weighted contribution to all ICs not in a Sector (x-axis) vs relative growth rate (y-axis), the Pearson's correlation is positive ($r = 0.33$, $p = 0.06$). (F) Plot showing the contribution to IC20 (x-axis) vs relative growth rate (y-axis), the Pearson's correlation is negative ($r = -0.35$, $p = 0.04$). It is possible that this correlation is an artifact of the strong negative impact on growth rate of mutating one position that strongly contributes to IC20. The majority of positions in the ribosome do not contribute strongly to IC20 and only 32 positions were assigned to IC20.

Tables S1-3 are available as downloadable csv files.

Table S1: Annotated list of Sector 1 residues.

Table S2: Annotated list of Sector 2 residues

Rows colored in pink are residues that are in the PTC. Rows colored in yellow are residues at the intersubunit interface. Rows colored in blue are residues in the exit tunnel. Rows colored in green are residues in the L1 stalk.

Table S3. Sector membership of A-minor motifs. A table showing the membership of bases involved in A-minor motifs identified in (12).

Table S4: Oligonucleotides. A list of oligonucleotides used in this study.

Oligonucleotide purpose	Oligonucleotide sequence
Forward primer for generating backbone for mutations 2056-2059 and 2064	CCGTGAACCTTTACTATAGCTTGACACTG
Reverse primer for generating backbone for mutations 2056-2059 and 2064	GTCTTGCCGCGGGTACACTG
Forward primer for generating backbone for mutations between 2112-2115	GAGGCTTTGAAGTGTGGACGCC
Reverse primer for generating backbone for mutations between 2112-2115	ACATCAAGGCTCAATGTTCAAGTGTCAAGC
Forward primer for generating backbone for mutations at 2249	TCTCCTCCTAAAGAGTAACGGAGG
Reverse primer for generating backbone for mutations at 2249	GTCAAAC TACCCACCAGACAC
Forward primer for generating backbone for mutations at 2465-2507	GGCTCATCACATCCTGGGGCTG
Reverse primer for generating backbone for mutations at 2465-2507	TCGATATGAACTCTTGGGCGGTATCAG
Forward primer for generating insert with mutation at 2056	GAAGATGCAGTGTACCCGCGGCAAGACG H GAAA
Forward primer for generating insert with mutation at 2057	GAAGATGCAGTGTACCCGCGGCAAGACG H GAAA
Forward primer for generating insert with mutation at 2058	GAAGATGCAGTGTACCCGCGGCAAGACGG B AA
Forward primer for generating insert with mutation at 2059	GAAGATGCAGTGTACCCGCGGCAAGACGG A BA
Forward primer for generating insert with mutation at 2064	GAAGATGCAGTGTACCCGCGGCAAGACGGAAA
Reverse primer for generating insert with mutations at 2056-2059	GTGTCAAGCTATAGTAAAGGTTACGGGGTC
Reverse primer for generating insert with mutation at 2064	GTGTCAAGCTATAGTAAAGGTTACGG H GTC
Forward primer for generating insert with mutation at 2112	GCTTGACACTGAACATTGAGCCTTGATGT H TAGGA
Forward primer for generating insert with mutation at 2113	GCTTGACACTGAACATTGAGCCTTGATGTG V AGGA

Forward primer for generating insert with mutation at 2114	GCTTGACACTGAACATTGAGCCTTGATGTGT B GGA
Forward primer for generating insert with mutation at 2115	GCTTGACACTGAACATTGAGCCTTGATGTGT A HGA
Reverse primer for generating insert with mutations at 2112-2115	GCAGACTGGCGTCCACACTTCAAAGCCTCCCACCTA
Forward primer for generating insert with mutation at 2249	GCGGACAGTGTCTGGTGGGTAGTTTGAC V GGG
Reverse primer for generating insert with mutation at 2249	GTGCTCCTCCGTTACTCTTTAGGAGGAGACCGC
Forward primer for generating insert with mutation at 2483	GGCTGATACCGCCCAAGAGTTCATATCGA D GGCGG
Forward primer for generating insert with mutation at 2484	GGCTGATACCGCCCAAGAGTTCATATCGAC H GGCGG
Forward primer for generating insert with mutation at 2485	GGCTGATACCGCCCAAGAGTTCATATCGAC H GGCGG
Forward primer for generating insert with mutation at 2486	GGCTGATACCGCCCAAGAGTTCATATCGAC D GGG
Forward primer for generating insert with mutations at 2497-2507	GGCTGATACCGCCCAAGAGTTCATATCGACGGCGG
Reverse primer for generating insert with mutations at 2483-2486	CCCCAGGATGTGATGAGCCGACATCGAGGTGCC
Reverse primer for generating insert with mutation at 2497	CCCCAGGATGTGATGAGCCGACATCGAG V GCC
Reverse primer for generating insert with mutation at 2499	CCCCAGGATGTGATGAGCCGACATCGA H GTGCC
Reverse primer for generating insert with mutation at 2502	CCCCAGGATGTGATGAGCCGACAT D GAGGTGCC
Reverse primer for generating insert with mutation at 2503	CCCCAGGATGTGATGAGCCGACA V CGAGGTGCC
Reverse primer for generating insert with mutation at 2505	CCCCAGGATGTGATGAGCCGA H ATCGAGGTGCC
Reverse primer for generating insert with mutation at 2507	CCCCAGGATGTGATGAGCC H ACATCGAGGTGCC
Forward primer to amplify region for sequencing	TGACTGGAGTTCAGACGTGtgctctccgatctNNNNNNCAGTGTAC CCGCGGCAAGAC
Reverse primer to amplify region for sequencing	CACTCTTTCCTACACGACgctctccgatctNNNNNNCCATACCCT TGGGACCTACTTCAGC
Forward barcode 1 for sequencing	CAAGCAGAAGACGGCATAACGAGATcgagtaatGTGACTGGAGTT CAGACGTG
Forward barcode 2 for sequencing	CAAGCAGAAGACGGCATAACGAGATtctccggaGTGACTGGAGTT CAGACGTG
Forward barcode 3 for sequencing	CAAGCAGAAGACGGCATAACGAGATaatgagcgGTGACTGGAGT TCAGACGTG
Reverse barcode 1 for sequencing	AATGATACGGCGACCACCGAGATCTACACtatagcctACACTCTT TCCCTACACGAC
Reverse barcode 2 for sequencing	AATGATACGGCGACCACCGAGATCTACACatagaggcACACTCT TCCCTACACGAC
Reverse barcode 3 for sequencing	AATGATACGGCGACCACCGAGATCTACACcctatcctACACTCTT TCCCTACACGAC
Reverse barcode 4 for sequencing	AATGATACGGCGACCACCGAGATCTACACggctctgaACACTCT TCCCTACACGAC

Reverse barcode 5 for sequencing	AATGATACGGCGACCACCGAGATCTACACcaggacgtACACTCT TTCCCTACACGAC
Reverse barcode 6 for sequencing	AATGATACGGCGACCACCGAGATCTACACgtactgacACACTCT TTCCCTACACGAC