LETTER

doi:10.1038/nature12543

Hidden specificity in an apparently nonspecific RNA-binding protein

Ulf-Peter Guenther[1,2], Lindsay E. Yandek[2], Courtney N. Niland[2], Frank E. Campbell[1], David Anderson[3], Vernon E. Anderson[2], Michael E. Harris[2] & Eckhard Jankowsky[1,2]

Nucleic-acid-binding proteins are generally viewed as either specific or nonspecific, depending on characteristics of their binding sites in DNA or RNA[1,2]. Most studies have focused on specific proteins, which identify cognate sites by binding with highest affinities to regions with defined signatures in sequence, structure or both[1–4]. Proteins that bind to sites devoid of defined sequence or structure signatures are considered nonspecific[1,2,5]. Substrate binding by these proteins is poorly understood, and it is not known to what extent seemingly nonspecific proteins discriminate between different binding sites, aside from those sequestered by nucleic acid structures[6]. Here we systematically examine substrate binding by the apparently nonspecific RNA-binding protein C5, and find clear discrimination between different binding site variants. C5 is the protein subunit of the transfer RNA processing ribonucleoprotein enzyme RNase P from *Escherichia coli*. The protein binds 5′ leaders of precursor tRNAs at a site without sequence or structure signatures. We measure functional binding of C5 to all possible sequence variants in its substrate binding site, using a high-throughput sequencing kinetics approach (HITS-KIN) that simultaneously follows processing of thousands of RNA species. C5 binds different substrate variants with affinities varying by orders of magnitude. The distribution of functional affinities of C5 for all substrate variants resembles affinity distributions of highly specific nucleic acid binding proteins. Unlike these specific proteins, C5 does not bind its physiological RNA targets with the highest affinity, but with affinities near the median of the distribution, a region that is not associated with a sequence signature. We delineate defined rules governing substrate recognition by C5, which reveal specificity that is hidden in cellular substrates for RNase P. Our findings suggest that apparently nonspecific and specific RNA-binding modes may not differ fundamentally, but represent distinct parts of common affinity distributions.

The term 'nonspecific' is widely used to describe proteins that bind DNA or RNA substrates at sites without apparent sequence or structure signatures[1,2,5]. Although nonspecific proteins are numerous and have many important biological roles, a key open question is whether the absence of defined recognition elements in nucleic-acid-binding sites reflects largely indiscriminate substrate binding, or whether and how nonspecific proteins discriminate between different binding sites. To answer this question, we systematically examined substrate binding for the apparently nonspecific RNA-binding protein C5, the protein subunit of RNase P from *E. coli*. RNase P is a ribonucleoprotein enzyme that removes 5′ leader sequences from precursor tRNA (ptRNA) in bacteria[7] (Fig. 1a). The C5 protein promotes ptRNA processing by RNase P[8], and contributes to ptRNA binding by associating with six consecutive nucleotides in the 5′ ptRNA leaders[9,10] (Fig. 1a, b). This binding site displays no apparent sequence or structure signatures in the 87 genomically encoded *E. coli* ptRNA leaders (Extended Data Fig. 1).

To determine whether and how C5 discriminates between different binding sites, we measured functional binding of C5 to all sequence variants in its cognate ptRNA site. Here, functional binding reflects

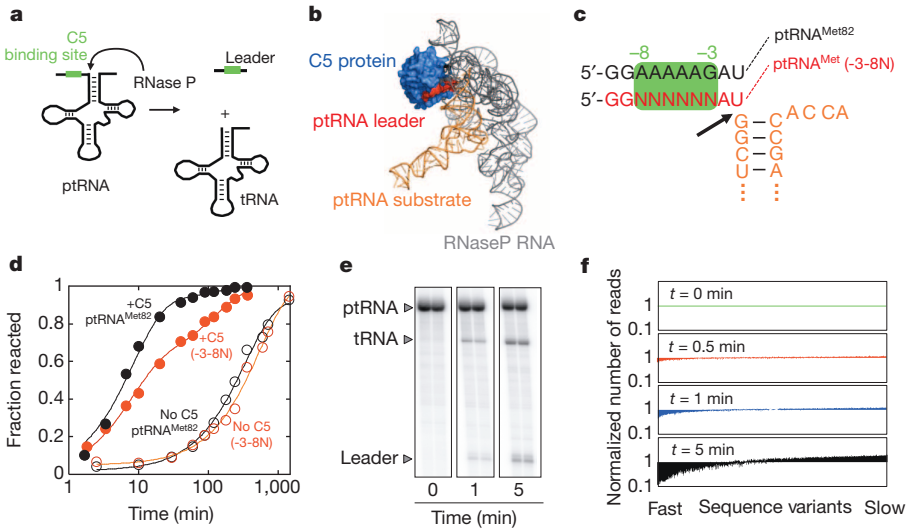


Figure 1 | Processing of precursor tRNA with randomized leader sequences. **a**, ptRNA processing reaction by RNase P. **b**, Structure of the RNase P holoenzyme[9]. **c**, Sequences of non-initiator ptRNA^{Met} leaders (reference, black; randomized, red). The tRNA body is omitted for clarity. The arrow indicates the cleavage site. **d**, Time courses of RNase P processing of ptRNA^{Met82} (black) and ptRNA$^{Met(-3-8N)}$ (red), in the presence (filled circles), and in the absence (open circles) of C5. The solid lines are fits to the integrated rate equation for a biphasic first order reaction. **e**, Polyacrylamide gel electrophroresis (PAGE) of reactions processed for Illumina sequencing. **f**, Distributions of species for individual time points, ranked from fastest to slowest. The *y* axis marks the change in read numbers for each substrate species at the reaction time indicated, normalized to the number of reads at $t = 0$. Colours emphasize the different reaction times.

[1]Center for RNA Molecular Biology, Case Western Reserve University, Cleveland, Ohio 44106, USA. [2]Department of Biochemistry, School of Medicine, Case Western Reserve University, Cleveland, Ohio 44106, USA. [3]Department of Management, Zicklin School of Business, Baruch College, The City University of New York, New York 10010, USA.

productive substrate association in an ongoing enzymatic reaction. It is expressed by the specificity constant (k_{cat}/K_m, the ratio of turnover number and Michaelis constant) for a given substrate variant, which measures biologically relevant specificity[11,12]. To determine functional binding of C5 to all substrate variants simultaneously, we generated non-initiator precursor tRNA^{Met} with a randomized C5-binding site (ptRNA$^{Met(-3\text{-}8N)}$, Fig. 1c), and followed the processing reaction of this substrate population (Fig. 1d). Reactions were conducted with excess ptRNA$^{Met(-3\text{-}8N)}$. Under these multiple turnover conditions all sequence variants compete for C5 association, and the relative reaction rate for each variant reflects functional binding[13].

The time course for the reaction of the randomized ptRNA$^{Met(-3\text{-}8N)}$ population differed markedly from the time course of ptRNA^{Met82} with a genomically encoded leader (Fig. 1d). This difference indicates that sequence variation affects functional binding by C5. Removal of C5 slowed the reaction rate as expected and greatly diminished the kinetic differences between the substrates with the genomically encoded and the randomized leaders (Fig. 1d).

To determine reaction rate constants for the individual substrate variants, we isolated remaining substrates at various reaction times and measured the distribution of the RNA species by Illumina sequencing (Fig. 1e, f, Extended Data Fig. 2 and Extended Data Table 1). We used primers with degenerate barcodes to detect biased amplification of sequences during the PCR (Extended Data Fig. 2 and Extended Data Table 1). Of the 4,096 sequence variants, 2,900 showed unbiased amplification and were retained for further analysis. The distribution of sequence variants changed over the reaction time, revealing distinct fast- and slow-reacting species (Fig. 1f). These data demonstrate that C5 discriminates between different sequence variants, despite the lack of sequence signatures in genomically encoded *E. coli* ptRNA leaders.

We calculated a relative processing rate constant (k^{rel}) for each RNA variant, using internal competition analysis, developed for the evaluation of kinetic isotope effects (Extended Data Fig. 3)[13–15]. The k^{rel} value is the ratio between the k_{cat}/K_m values for the given sequence variant and our reference sequence, the physiological leader AAAAAG. The relative rate constants for all sequence variants describe C5 binding to the entire sequence space of the six-nucleotide recognition site. Our approach to measure functional binding of large numbers of substrates during an ongoing reaction adds a kinetic dimension to the scope of high-throughput sequencing experiments with randomized RNA populations[3,4,16,17]. We therefore propose to term our method high-throughput sequencing kinetics (HITS-KIN). The approach is applicable to other systems for kinetic analysis of next generation sequencing data.

For the ptRNA processing reaction with C5, the HITS-KIN method revealed a range of relative rate constants spanning several orders of magnitude (Fig. 2a). Obtained relative rate constants were highly reproducible in independent experiments (Fig. 2b). We also validated rate constants by direct kinetic measurements of selected sequence variants (Fig. 2c and Extended Data Fig. 4). Together, these data show that the HITS-KIN approach provides reproducible and accurate relative rate constants.

We next plotted the number of sequence variants processed at a given range of relative rate constants (Fig. 2d). The resulting histogram revealed that a significant number of sequence variants reacted faster than the physiological leader reference ($k^{rel} > 1$). Numerous sequence variants reacted slower ($k^{rel} < 1$). These observations indicate that physiological leader sequences of non-initiator ptRNA^{Met} are not preferentially bound by C5. Removal of C5 greatly contracted the range of relative rate constants, highlighting the impact of C5 on functional substrate binding and on the characteristic affinity distribution (Extended Data Fig. 5).

Most notably, the shape of the distribution of functional C5 affinities closely resembled affinity distributions of highly specific DNA-binding proteins, for which large numbers of sequence variants had been examined[18–21] (Fig. 2d). This degree of similarity between the non-specific C5 and specific proteins was unexpected, given the absence of sequence signatures in the C5 binding site. For specific proteins, the

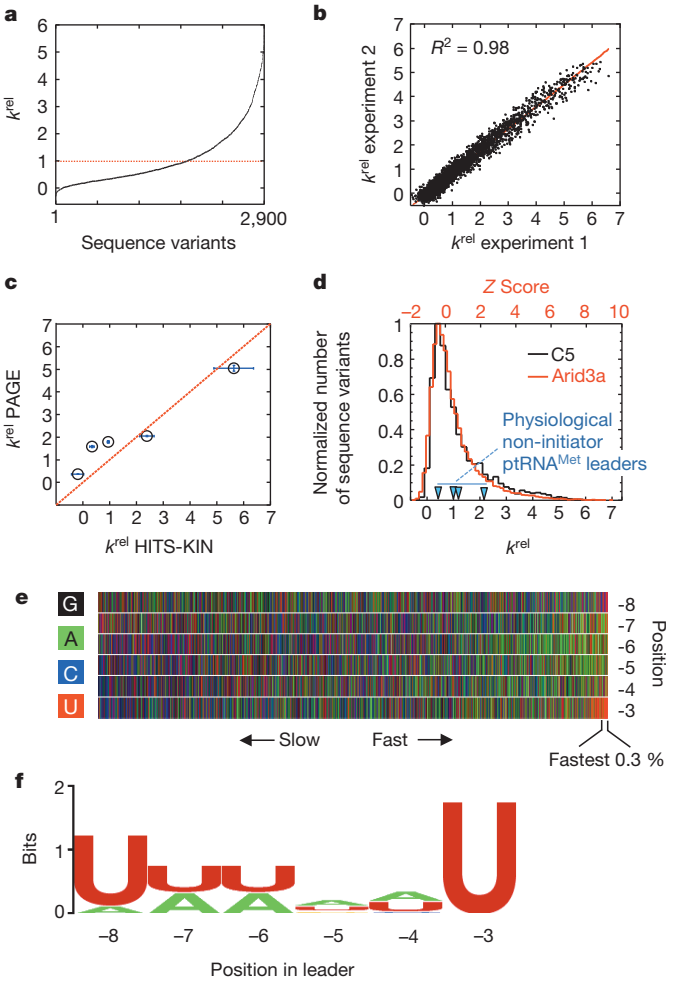


Figure 2 | Discrimination of C5 between different precursor tRNA^{Met} leader sequences. **a**, Relative rate constants (k^{rel}) for processing of all ptRNA leader sequence variants, ranked from slow to fast. Relative rate constants are averaged from four values (two time points of two experiments) and shown for only sequences where data from all four measurements passed quality control criteria (Extended Data Table 1). The line at $k^{rel} = 1$ marks the reference sequence. **b**, Correlation of relative rate constants from two independent biological replicates (red line, linear fit through the data; R^2, correlation coefficient). **c**, Correlation between relative rate constants obtained by PAGE and by the HITS-KIN approach for selected sequence variants. Error bars represent the s.d. of three or more individual measurements. **d**, Distribution of relative rate constants for processing of ptRNA$^{Met(-3\text{-}8N)}$ sequence variants by C5 (black) and apparent affinities for DNA binding by the transcription factor Arid3a, indicated as *Z*-scores based on published microarray data[18]. The *Z*-score is not identical to k^{rel} values, but accurately reflects affinity-based ranking of all sequences[18] (triangles, k^{rel} values for genomic leader sequences of ptRNA^{Met}). **e**, Plot of all sequence variants ranked from slowest to fastest processed. The bracket marks 0.3% of sequence variants with the largest relative rate constants. **f**, Sequence logo for this fraction.

cellular substrates that define binding site signatures are found at the high-affinity tail of the distribution[18,19] (Extended Data Fig. 6a, b). Remarkably, this high-affinity region for C5 also shows a clear sequence signature (Fig. 2e, f), as seen for specific proteins. In stark contrast to specific proteins, the C5 sequence signature does not correspond to the physiological binding sites on the non-initiator ptRNA^{Met}. None of the genomically encoded non-initiator ptRNA^{Met} leader sequences falls into this fastest-reacting fraction (Fig. 2d). For both C5 and specific proteins, no sequence signatures were detected for other regions of the sequence spectrum (Extended Data Fig. 6). Our results therefore reveal remarkable similarities between sequence discrimination by the apparently nonspecific C5 and by specific DNA-binding proteins. At the same time, our data

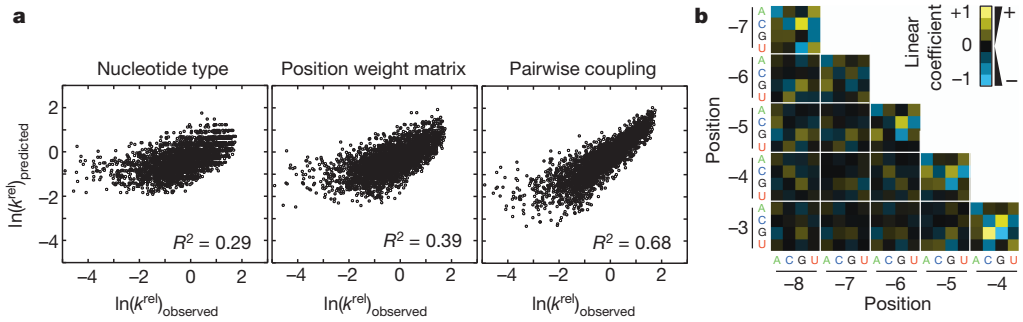


Figure 3 | Rules for sequence discrimination by C5. a, Correlation between observed k^{rel} and values calculated with the best fit of the data to models of increasing complexity. Logarithmic k^{rel} values are used because of their correspondence to differences in binding energies[30]. R^2 expresses the correlation of each model with measured processing rate constants. **b**, Functional coupling between two base positions. Yellow squares show promotion of processing (high linear coefficients), black squares indicate small or no effects, blue squares mark inhibition of processing.

highlight a major difference: sequences bound with the highest affinity do not represent physiological substrates for C5, but for specific DNA-binding proteins with known affinity distributions.

To delineate sequence determinants that govern substrate recognition by C5, we fit the distribution of rate constants to models of increasing complexity and determined which percentage of the measured variance in the rate constants was explained by the respective model. Our simplest model considered only the number of a given nucleotide in the binding site, regardless of position. This model explained 29% of the variance in the measured rate constants (Fig. 3a, left). The model suggested favourable binding of sequences rich in adenine and uracil (Extended Data Fig. 7a). As A–U base pairs are thermodynamically less stable than G–C base pairs, we speculate that the variance explained by this model reflects in part the propensity of the leader to form transient structures with other parts of the ptRNA[22], which potentially compete with C5 binding. Although competing structures are generally expected in RNAs with more than two dozen nucleotides[23], the relatively low correlation of the model with measured rate constants suggests that competing RNA structures have only limited impact on C5 binding for the majority of sequences.

We next considered both base identity and position in the binding site. This model, a traditional position weight matrix[21], explained 39% of the variance in measured rate constants (Fig. 3a, middle, and Extended Data Fig. 7b). This modest improvement over the previous model indicated that the position of individual bases in the binding site impacted C5 binding only to a limited extent. However, the position weight matrix assesses the bases independently of each other[21]. To probe inter-dependence of the bases in the binding site, we used a model accounting for functional coupling between two bases. This model explained 68% of the variance in measured rate constants (Fig. 3a, right). The strongest couplings were detected between neighbouring bases (Fig. 3b).

The observed strength of the couplings between adjacent bases did not scale with energies expected to overcome stacking of the respective bases[22]. This finding suggests that the couplings result from interactions of the RNA with C5, not primarily from inherent RNA conformations. Functional couplings between more than two base positions, assessed by neural network analysis, only modestly improved correlation between predicted and measured data, and explained 76% of the variance (Extended Data Fig. 8). Thus, functional couplings between adjacent bases exert the largest influence on C5 binding. The limited resolution of the structural model of RNase P protein bound to RNA[9] currently precludes structural interpretation of these effects. However, we note that functional coupling between neighbouring bases also contributes markedly to the binding of several specific transcription factors to DNA[21,24,25].

Taken together, the examination of the functional binding data with models of increasing complexity reveals defined rules for substrate binding by C5. The data demonstrate that discrimination between different substrates, and thus specificity, is an inherent property of C5. However, this specificity is 'hidden' in the cellular RNA targets. This observation raises the question of why the specificity in C5 has not led to selection of ptRNA leaders with high-affinity sequence signatures, as seen in proteins with canonical specificity[18–21]. Our data suggest a further-reaching utility of specificity. C5 uses its inherent specificity, as reflected in the rules for substrate recognition, to enable binding of diverse substrate variants with similar functional affinity. This enables RNase P to process these diverse substrates at a similar rate, which may be required for cellular tRNA homeostasis[26].

The marked similarities between affinity distributions of C5 and those of highly specific transcription factors also raise questions about the concept of 'nonspecific' RNA-binding proteins. Given that RNA binding requires a protein interface to establish interactions with the RNA, certain RNA sequence or structure variants conceivably fit this interface better than others. Genuine nonspecificity may therefore be difficult to accomplish, even for proteins binding exclusively to the RNA backbone, because sequence differences impact backbone geometry[27]. Differences between substrate variants may become smaller for proteins that bind to the backbone of RNA duplexes, which show less structural heterogeneity, but are nevertheless dynamic[28].

Preferences of apparently nonspecific proteins for certain binding-site variants are thus likely to impact substrate selection, unless compensation for these preferences exist. Compensation may arise from varying concentrations of RNA species, rate-determining metabolic steps other than substrate binding, or a combination of these. Alternatively, a single protein could bind multiple distinct substrate regions while thermodynamically compensating for the preferences at each region, as shown for uniform binding of diverse aminoacyl-tRNAs to elongation factor Tu[29].

Although hidden specificity remains to be revealed for other proteins, the findings for C5 indicate that absence of sequence or structure signatures in cellular binding sites does not reflect an inability to discriminate between different RNA binding sites. At the same time, the data highlight the key difference between the hidden specificity of C5 and proteins that are specific in a canonical sense. For proteins with canonical specificity, cellular substrates seem to fall mainly into the high-affinity region of the sequence distribution. This region is associated with sequence signatures, even for C5. Biological substrates for C5 bind near the median of the affinity distribution, which does not produce a sequence signature. These findings suggest that specific and nonspecific binding modes may not fundamentally differ, but represent distinct parts of similar affinity distributions. Our data therefore have potentially broad implications for RNA binding by proteins thought to be nonspecific, including many RNases, RNA helicases or the La-protein.

METHODS SUMMARY

ptRNAs and ptRNA^{Met} with randomized leader sequences were produced by *in vitro* transcription from PCR-generated templates. RNase P processing reactions were carried out with 1 μM ptRNA and 5 nM RNase P holoenzyme (equimolar RNase P RNA and C5). Product and unreacted ptRNA were separated by PAGE. Complementary

DNA libraries for Illumina sequencing were prepared from unreacted ptRNA at each given time point. Primers with degenerate barcodes were used to detect biased PCR amplification of certain sequences. Sequencing was performed on an Illumina GA2. Relative rate constants k^{rel} for individual substrate variants were calculated from changes in the distribution of substrates over time, using a multiple turnover reaction scheme for competitive substrate kinetics, which was extended to several thousand substrates. Computational modelling for the rules of substrate discrimination was performed by ordinary least squares regression of the matrix of values for $\ln(k^{rel})$ for each sequence variant according to four models of increasing complexity. The quality of the different models was judged by the correlation coefficient between a data set calculated from values obtained from the regression analysis and the set of experimentally obtained values for $\ln(k^{rel})$.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 15 May; accepted 14 August 2013.
Published online 22 September 2013.

1. Gupta, A. & Gribskov, M. The role of RNA sequence and structure in RNA–protein interactions. *J. Mol. Biol.* **409,** 574–587 (2011).
2. von Hippel, P. H. & Berg, O. G. On the specificity of DNA–protein interactions. *Proc. Natl Acad. Sci. USA* **83,** 1608–1612 (1986).
3. Ray, D. *et al.* Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nature Biotechnol.* **27,** 667–670 (2009).
4. Campbell, Z. T. *et al.* Cooperativity in RNA–protein interactions: global analysis of RNA binding specificity. *Cell Rep.* **1,** 570–581 (2012).
5. Singh, R. & Valcárcel, J. Building specificity with nonspecific RNA-binding proteins. *Nature Struct. Mol. Biol.* **12,** 645–653 (2005).
6. Zhuang, F., Fuchs, R. T., Sun, Z., Zheng, Y. & Robb, G. B. Structural bias in T4 RNA ligase-mediated 3′-adapter ligation. *Nucleic Acids Res.* **40,** e54 (2012).
7. Kurz, J. C. & Fierke, C. A. Ribonuclease P: a ribonucleoprotein enzyme. *Curr. Opin. Chem. Biol.* **4,** 553–558 (2000).
8. Smith, J. K., Hsieh, J. & Fierke, C. A. Importance of RNA–protein interactions in bacterial ribonuclease P structure and catalysis. *Biopolymers* **87,** 329–338 (2007).
9. Reiter, N. J. *et al.* Structure of a bacterial ribonuclease P holoenzyme in complex with tRNA. *Nature* **468,** 784–789 (2010).
10. Rueda, D., Hsieh, J., Day-Storms, J. J., Fierke, C. A. & Walter, N. G. The 5′ leader of precursor $tRNA^{Asp}$ bound to the *Bacillus subtilis* RNase P holoenzyme has an extended conformation. *Biochemistry* **44,** 16130–16139 (2005).
11. Herschlag, D. The role of induced fit and conformational changes of enzymes in specificity and catalysis. *Bioorg. Chem.* **16,** 62–96 (1988).
12. Fersht, A. R. *Enzyme Structure and Mechanism* (Freeman, 1985).
13. Cornish-Bowden, A. Enzyme specificity: its meaning in the general case. *J. Theor. Biol.* **108,** 451–457 (1984).
14. Cleland, W. W. in *Isotope Effects in Chemistry and Biology* (eds Kohen, A. & Limbach, H. H.) 915–930 (CRC Press, 2006).
15. Schellenberger, V., Siegel, R. A. & Rutter, W. J. Analysis of enzyme specificity by multiple substrate kinetics. *Biochemistry* **32,** 4344–4348 (1993).
16. Lorenz, C. *et al.* Genomic SELEX for Hfq-binding RNAs identifies genomic aptamers predominantly in antisense transcripts. *Nucleic Acids Res.* **38,** 3794–3808 (2010).
17. Pitt, J. N. & Ferré-D'Amaré, A. R. Rapid construction of empirical RNA fitness landscapes. *Science* **330,** 376–379 (2010).
18. Badis, G. *et al.* Diversity and complexity in DNA recognition by transcription factors. *Science* **324,** 1720–1723 (2009).
19. Rowe, W. *et al.* Analysis of a complete DNA–protein affinity landscape. *J. R. Soc. Interface* **7,** 397–408 (2010).
20. Nutiu, R. *et al.* Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument. *Nature Biotechnol.* **29,** 659–664 (2011).
21. Stormo, G. D. & Zhao, Y. Determining the specificity of protein–DNA interactions. *Nature Rev. Genet.* **11,** 751–760 (2010).
22. SantaLucia J. Jr & Turner, D. H. Measuring the thermodynamics of RNA secondary structure formation. *Biopolymers* **44,** 309–319 (1997).
23. Forsdyke, D. R. Calculation of folding energies of single-stranded nucleic acid sequences: conceptual issues. *J. Theor. Biol.* **248,** 745–753 (2007).
24. Maerkl, S. J. & Quake, S. R. A systems approach to measuring the binding energy landscapes of transcription factors. *Science* **315,** 233–237 (2007).
25. Zhao, Y. & Stormo, G. Quantitative analysis demonstrates most transcription factors require only simple models of specificity. *Nature Biotechnol.* **29,** 480–483 (2011).
26. Sun, L., Campbell, F. E., Yandek, L. E. & Harris, M. E. Binding of C5 protein to P RNA enhances the rate constant for catalysis for P RNA processing of pre-tRNAs lacking a consensus (+ 1)/C(+ 72) pair. *J. Mol. Biol.* **395,** 1019–1037 (2010).
27. Leontis, N. B., Lescoute, A. & Westhof, E. The building blocks and motifs of RNA architecture. *Curr. Opin. Struct. Biol.* **16,** 279–287 (2006).
28. Snoussi, K. & Leroy, J. L. Imino proton exchange and base-pair kinetics in RNA duplexes. *Biochemistry* **40,** 8898–8904 (2001).
29. LaRiviere, F. J., Wolfson, A. D. & Uhlenbeck, O. C. Uniform binding of aminoacyl-tRNAs to elongation factor Tu by thermodynamic compensation. *Science* **294,** 165–168 (2001).
30. Stormo, G. D., Schneider, T. D. & Gold, L. Quantitative analysis of the relationship between nucleotide sequence and functional activity. *Nucleic Acids Res.* **14,** 6661–6679 (1986).

Supplementary Information is available in the online version of the paper.

Acknowledgements We particularly thank T. Nilsen for comments on the manuscript. We are grateful to G. Stormo for discussion, M. Adams for support with the Illumina sequencing, and H.-C. Lin for technical assistance. This work was supported by the US National Institutes of Health (NIH) (GM067700, GM099720 and CSTA UL1RR024989 to E.J.; GM056740 and GM096000 to M.E.H.; T32 GM008056 to C.N.N.). U.-P.G. received a DFG fellowship.

Author Contributions U.-P.G., M.E.H. and E.J. designed the study. U.-P.G., L.E.Y., C.N.N. and F.E.C. performed the experiments. V.E.A. contributed to the development of the data analysis framework. D.A. developed and performed the modelling for binding models. U.-P.G., D.A., M.E.H. and E.J. analysed the data. U.P.G., M.E.H. and E.J. wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.E.H. (meh2@case.edu) or E.J. (exj13@case.edu).

METHODS

E. coli RNase P holoenzyme and RNase P RNA were prepared and tested for integrity as described previously[31,32]. The non-initiator ptRNA^{Met} substrates contain 8 nucleotides of the genomically encoded leader (Fig. 1c), and 21 nucleotides at the 5′ end for Illumina sequencing (Extended Data Fig. 2). These RNAs were generated by *in vitro* transcription from DNA generated by PCR amplification of the ptRNA^{Met82} gene (*PMET82*). The forward primer introduced the T7 promoter sequence and the additional 21 nucleotides (Extended Data Fig. 4). The ptRNA$^{Met(-3-8N)}$ substrate population with randomized leader sequence N(−3) to N(−8) was generated using a primer with this region randomized (NNNNNN).

The following PCR primers were used (C5 binding site is underlined):

ptRNAMet82F, 5′-TAATACGACTCACTATAGGGAGACCGGAATTCAGATTGATGAAAAAGATGGCTACGTAGCTCAGTTGG-3′; ptRNA^{Met82}F$_{Eco}$, 5′-GGGTTAACCTAATACGACTCACTATAGGGAGACCGGAATTCAGATTGATGAAAAAGATGGCTACGTAGCTCAGTTGG-3′; ptRNA^{Met}F$_{randomized}$, 5′-TAATACGACTCACTATAGGGAGACCGGAATTCAGATTGATGNNNNNNATGGCTACGTAGCTCAGTTGG-3′; ptRNA^{Met82}R, 5′-TGGTGGCTACGACGGGATTC-3′; ptRNA^{Met82}R$_{Bbs}$, 5′-CGGGATCCGAAGACAGTGGTGGCTACGACGGGATTC-3′. DNA templates for substrates L1 to L5 (Fig. 2c) contained the following C5-binding sites: L1, TTATAT; L2, TCAGAC; L3, ATTCAA; L4, CGTCAG; L5, CTCCTG.

PCR protocol: 95 °C, 2 min; 30 cycles (95 °C, 30 s at 55 °C, 45 s at 72 °C), final extension at 72 °C for 5 min.

The PCR products (142 base pairs (bp)) were extracted with phenol and chloroform and recovered by ethanol precipitation. PCR products for the ptRNA^{Met82} DNA were amplified with the ptRNA^{Met82}F$_{Eco}$ and ptRNA^{Met82}RBbs primers, which include BamHI and EcoRI restriction sites. The PCR product was digested with these enzymes and cloned into pUC19. The resulting plasmid, pPTRNAmetT(+21), was digested with BbsI to yield the template for *in vitro* transcription with the correct ptRNA^{Met82} 3′ end.

In vitro transcription was performed in a volume of 400 µl with 15–20 µg of PCR template or cloned plasmid DNA template, 400 enzyme units of T7 RNA polymerase (Ambion), 0.01 unit yeast pyrophosphatase, 0.5 mM NTP, and the reaction buffer supplied by the polymerase manufacturer was supplemented with 2.5 mM $MgCl_2$. Reactions were incubated overnight at 37 °C. The full-length RNA was purified on 8% denaturing PAGE, as described previously[31,32].

Recovered ptRNAs were dephosphorylated using calf intestinal phosphatase and 5′ end labelled with ^{32}P using γ^{32}P-ATP and T4 polynucleotide kinase according to standard methods. For the HITS-KIN experiments, the RNA was uniformly labelled with γ^{32}P-GTP in the *in vitro* transcription (NTPs 100 µM).

RNase P processing reactions. Multiple turnover reactions were performed in buffer containing 50 mM Tris-HCl, pH 8.0, 100 mM NaCl, 17.5 mM $MgCl_2$, 0.005% Triton X-100, with 1 µM ptRNA and 5 nM *E. coli* RNase P holoenzyme (1:1 ratio of P RNA and C5 protein). Equal volumes (40 µl) of enzyme and radiolabelled substrate at two times their final concentrations were prepared in reaction buffer and combined to initiate the reaction. Aliquots (5 µl) were removed at the times indicated for 5% to 30% substrate conversion. The reactions were quenched by addition of a solution (5 µl) containing formamide and 100 mM EDTA. ptRNA and reaction products were resolved on 10% denaturing PAGE (Fig. 1e). The fraction substrate converted to product was determined with a PhosphorImager (GE) and the ImageQuant software. Subsequently, precursor bands in the gel were located by exposure to X-ray film. The bands were excised and eluted as described previously[32]. Eluted RNA was extracted with phenol and chloroform, and recovered by ethanol precipitation.

Relative rate constants for individual non-initiator ptRNA^{Met} substrates (L1–L5, Fig. 2c, for defined sequences of C5 binding site, see above) were determined in reactions containing 1 µM of the pool of randomized ptRNA$^{Met(-3-8N)}$, spiked with trace amounts (<0.1 nM) of the respective radiolabelled L1–L5 substrate. Time courses of the reactions were followed as described above and apparent rate constants were determined from plots of product accumulation over time[32]. As outlined below, the ratio of the observed rate constants is k^{rel} because in competition kinetics the substrates at the concentrations used behave as *V*/*K* systems. *V*/*K* is proportional to k_{cat}/K_m at a given substrate concentration.

DNA library preparation. Complementary DNA libraries for Illumina sequencing were prepared from unreacted ptRNA, recovered from PAGE as described above. RNA was resuspended in 25 µl H_2O, and concentration was determined with a Beckman ultraviolet spectrophotometer. First-strand synthesis was performed with 4 pmol of this RNA in a 20-µl standard reaction containing 1 µM reverse transcription primer (Extended Data Fig. 2a) and 0.5 µl Superscript III (Invitrogen) for 10 min at 42 °C, 40 min at 50 °C and 20 min at 55 °C. The reaction was stopped by incubation at 95 °C for 5 min. The generated cDNA was diluted (1:300). One microlitre of this solution was used in PCR reactions with 1.25 enzyme units Herculase polymerase (Stratagene), reverse transcription primer (0.5 µM) and indexed forward primer (0.5 µM) for 2 min at 98 °C, followed by 19 cycles (15 s at 98 °C, 20 s at 59 °C, 20 s at 72 °C), and incubation for 10 min at 72 °C. PCR products were purified with P6 microcentrifuge columns (Bio-Rad) and analysed by agarose gel electrophoresis (Extended Data Fig. 2b). The solutions were pooled in an equimolar fashion and sequenced in a single lane of an Illumina GA2, according to the manufacturer's protocols. Primer sequences were as follows: reverse transcription primer, 5′-CAAGCAGAAGACGGCATACGATGGTGGCTACGACGGGAT-3′; indexed forward primers (NN, degenerate barcode; underlined letters, index barcode), 5′-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNATCGGGAGACCGGAATTCAGATTG-3′; 5′-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNGATGGGAGACCGGAATTCAGATTG-3′; 5′-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNCGAGGGAGACCGGAATTCAGATTG-3′; 5′-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNTCCGGGAGACCGGAATTCAGATTG-3′.

Processing of Illumina sequencing data. All reads were aligned to the sequence of nucleotides 6–29 of the read (Extended Data Fig. 2c), permitting one mismatch but no gaps, using the basic local alignment search tool (BLAST). Aligned reads were then sorted according to their index tag, and separated into different files. For corresponding statistics, see Extended Data Table 1.

We probed possible over- or under-amplification of certain sequences during the PCR using the two-nucleotide degenerate barcode (positions 1 and 2, Extended Data Fig. 2c). Correctly amplified sequences show a distribution of degenerated barcode nucleotide combinations that is identical, within error, for all leader species. Both over- and under-amplification cause deviations from this distribution. We determined the distribution of all nucleotide combinations at positions 1 and 2 for each leader sequence. The expected distribution of the two-nucleotide degenerate barcode was calculated from all 4,096 leader sequence variants. Then, chi-squared tests were performed for each leader variant. Sequences for which the threshold exceeded $\alpha > 0.05$ were excluded from further analysis (between 4% and 10% of the sequence variants, Extended Data Table 1).

Determination of relative rate constants k^{rel} from Illumina sequencing data. Rate constants for individual substrate variants were calculated from time-dependent changes of the distribution of substrate variants (Fig. 1f), using a multiple turnover reaction scheme for competitive alternative substrate kinetics[15,33] (Extended Data Fig. 3).

The observed rate constant ($v_{1...i}$) for processing of one individual substrate (S_1) is proportional to the fraction of total enzyme that binds this substrate to form a complex (ES_1) that further reacts to form product and regenerates free enzyme according to:

$$v_1 = V_1 E f_{ES1} \tag{1}$$

Here, V_1 is the first order rate constant for the reaction of ES_1 to yield free product (Extended Data Fig. 3), f_{ES1} is the fraction of total (active) enzyme (E) in the ES_1 complex. Additional substrates act essentially as competitive inhibitors of the multiple turnover reaction. Accordingly, v_1 is:

$$v_1 = \frac{V_1 E \frac{S_1}{K_1}}{\left(1 + \sum_{i=1}^{n} \frac{S_i}{K_i}\right)} \tag{2}$$

Variables are defined in Extended Data Fig. 3. By extension, v_2 is:

$$v_2 = \frac{V_2 E \frac{S_2}{K_2}}{\left(1 + \sum_{i=1}^{n} \frac{S_i}{K_i}\right)} \tag{3}$$

As the denominators of equations (2) and (3) are the same, the ratio of two observed rate constants (v_1/v_2) therefore becomes:

$$\frac{v_2}{v_1} = \frac{\left(\frac{V}{K}\right)_2}{\left(\frac{V}{K}\right)_1} \left(\frac{S_2}{S_1}\right) \tag{4}$$

We define the parameter k^{rel} as the ratio of the *V*/*K* values of a given substrate (S_2) to a reference substrate (S_1):

$$\frac{\left(\frac{V}{K}\right)_2}{\left(\frac{V}{K}\right)_1} = k^{rel} \tag{5}$$

and thus:

$$\frac{v_2}{v_1} = k^{\mathrm{rel}}\left(\frac{S_2}{S_1}\right) \tag{6}$$

The reference substrate S_1 is the genomically encoded leader sequence for the ptRNA$^{\mathrm{Met82}}$ (AAAAAG)[34]. Thus, $k^{\mathrm{rel}} > 1$ for a ptRNA variant that reacts faster than the reference substrate ($V_i/K_i > V_1/K_1$), whereas $k^{\mathrm{rel}} < 1$ indicates a slower reaction ($V_i/K_i < V_1/K_1$).

Equations (4) to (6) highlight three important points regarding the use of internal competition kinetics for the analysis of deep sequencing data. First, both substrates will behave as V/K systems[35,36] regardless of the substrate concentrations. This is true even if one or both concentrations are greater than the respective values for K_{m}, because both substrates compete for free enzyme[13,15]. Second, the ratio of observed rate constants and the ratio of V/K values are independent of enzyme concentration, provided the steady state conditions are maintained. Third, the reaction step that limits V/K does not have to be the same for both substrates.

Integration of equation (5) over time ensures validity of the expression for any reaction interval[13], and we obtain:

$$k^{\mathrm{rel}} = \frac{\ln\left(\frac{S_2}{S_{2,0}}\right)}{\ln\left(\frac{S_1}{S_{1,0}}\right)} \tag{7}$$

Here, $S_{1,0}$ and $S_{2,0}$ are the initial concentrations of the two substrates. S_1 (reference substrate) and S_2 (the specific sequence variant) are the respective concentrations at a defined time interval. The quantities that can be measured are the relative concentrations of S_2 and S_1; that is, S_2/S_1 and $S_{2,0}/S_{1,0}$. We define these quantities as the ratios (R) between substrates:

$$R_i = \frac{S_i}{S_1} \tag{8}$$

$$R_{i,0} = \frac{S_{i,0}}{S_{1,0}} \tag{9}$$

The initial mole fractions (X_i) of S_i are defined as:

$$X_i = \frac{S_i}{\sum_{i=1}^{n} S_{i,0}} \tag{10}$$

$S_{\mathrm{i,0}}$ is the concentration of a given substrate at the reaction start, S_{i} is the concentration at a time point where the overall reaction amplitude for the entire substrate population has reached the value f. We obtain:

$$f = 1 - \sum_{i=1}^{n} X_i \tag{11}$$

Analogous to the treatment of kinetic isotope effects using internal competition in a previous publication[14], we insert the defined mole fractions and substrate ratios (equations (8) to (11)) into equation (11). This is rearranged, and the result is the following equation:

$$\frac{S_i}{S_{i,0}} = \frac{(1-f)}{\frac{R_{i,0}}{R_i}\sum_{i=1}^{n}\left(\frac{R_i}{R_{i,0}}X_i\right)} \tag{12}$$

We substitute this term in equation (7), and consider that substrate ratios at time zero equal one.

$$\frac{R_1}{R_{1,0}} = 1 \tag{13}$$

We obtain the following expression for the relative rate constant for any substrate, S_i:

$$k_i^{\mathrm{rel}} = \frac{\ln\frac{(1-f)}{\frac{R_{i,0}}{R_i}\left(\sum_1^i \frac{R}{R_0}X\right)}}{\ln\frac{(1-f)}{\sum_1^i \frac{R}{R_0}X}} \tag{14}$$

Here, R is the ratio of each sequence (including S_1) to S_1, R_0 is the ratio of each variant to $S_{1,0}$ at the reaction start, and X is the mole fraction for a given sequence variant. The method outlined above is applicable to any technique capable of determining substrate ratios (for example, mass spectrometry, isotopic counting, chromatography).

We computed R and X values for each substrate using the filtered number of counts for each variant, obtained from Illumina sequencing (Supplementary Table 2). The overall fraction of reacted product was determined by PhosphorImager analysis of the PAGE (Fig. 1e).

In principle, values for k^{rel} can be computed at any value of f. However, there is little relative change in the number of sequencing reads at early time points. However, at early time points the highest resolution is seen for the fastest reacting variants, while k^{rel} values for slower sequences are optimally measured at greater values of f. Values of $f = 0.1$ to 0.3 provided reliable measurements for most sequence variants. Nevertheless, for slow-reacting variants small changes in the number of sequencing reads at early time points are occasionally exceeded by sampling error in the number of reads, resulting in negative values for k^{rel}.

Computational modelling of rules for substrate discrimination. With regard to nucleotide type, this model considers the number of each nucleobase in the binding site, regardless of its position. For each sequence variant the corresponding value of $\ln(k^{\mathrm{rel}})$ (Fig. 3a) is described by a set of linear coefficients (β), according to the equation:

$$\ln\left(k^{\mathrm{rel}}\right) = \beta_0 + \beta_A \bullet A + \beta_C \bullet C + \beta_G \bullet G \tag{15}$$

A, C and G are the number of the respective nucleobases (explanatory value). The number of U follows from these variables and is therefore not included ($\beta_U = 0$). Equation (15) describes the average increase in $\ln(k^{\mathrm{rel}})$ corresponding to a one-unit increase in the explanatory variable. For example, for each additional C in the sequence, the $\ln(k^{\mathrm{rel}})$ increases by β_C.

Linear coefficients for the entire data set were computed by ordinary least squares (OLS) regression, using the open-source statistical package R (http://www.r-project.org/) and the exact equation:

$$\beta_{\mathrm{N}} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y \tag{16}$$

Here, Y is the vector of outcomes $\ln(k^{\mathrm{rel}})$ and X is the matrix of explanatory variables (A,C,G). X^{T} is X transposed, and X^{-1} is the inverse of X.

To compare predicted and measured $\ln(k^{\mathrm{rel}})$ values for all sequence variants (Fig. 3a), we calculated predicted values using equation (15), plotted these versus the corresponding measured value, and determined the correlation between measured and calculated data set. The coefficient of correlation (R^2) was computed according to[37]:

$$R^2 = 1 - \frac{S^2_{\mathrm{error}}}{S^2_{\mathrm{total}}} \tag{17}$$

S^2_{error} is the sum of squared errors and measures the error, or unexplained variance, in the regression. The error is the distance from each point to the regression line, and is calculated for each data point, squared, and summed, according to:

$$S^2_{\mathrm{error}} = \sum_i (y_i - f_i)^2 \tag{18}$$

S^2_{total} is the sample variance, which is calculated according to:

$$S^2_{\mathrm{total}} = \sum_i (y - \bar{y})^2 \tag{19}$$

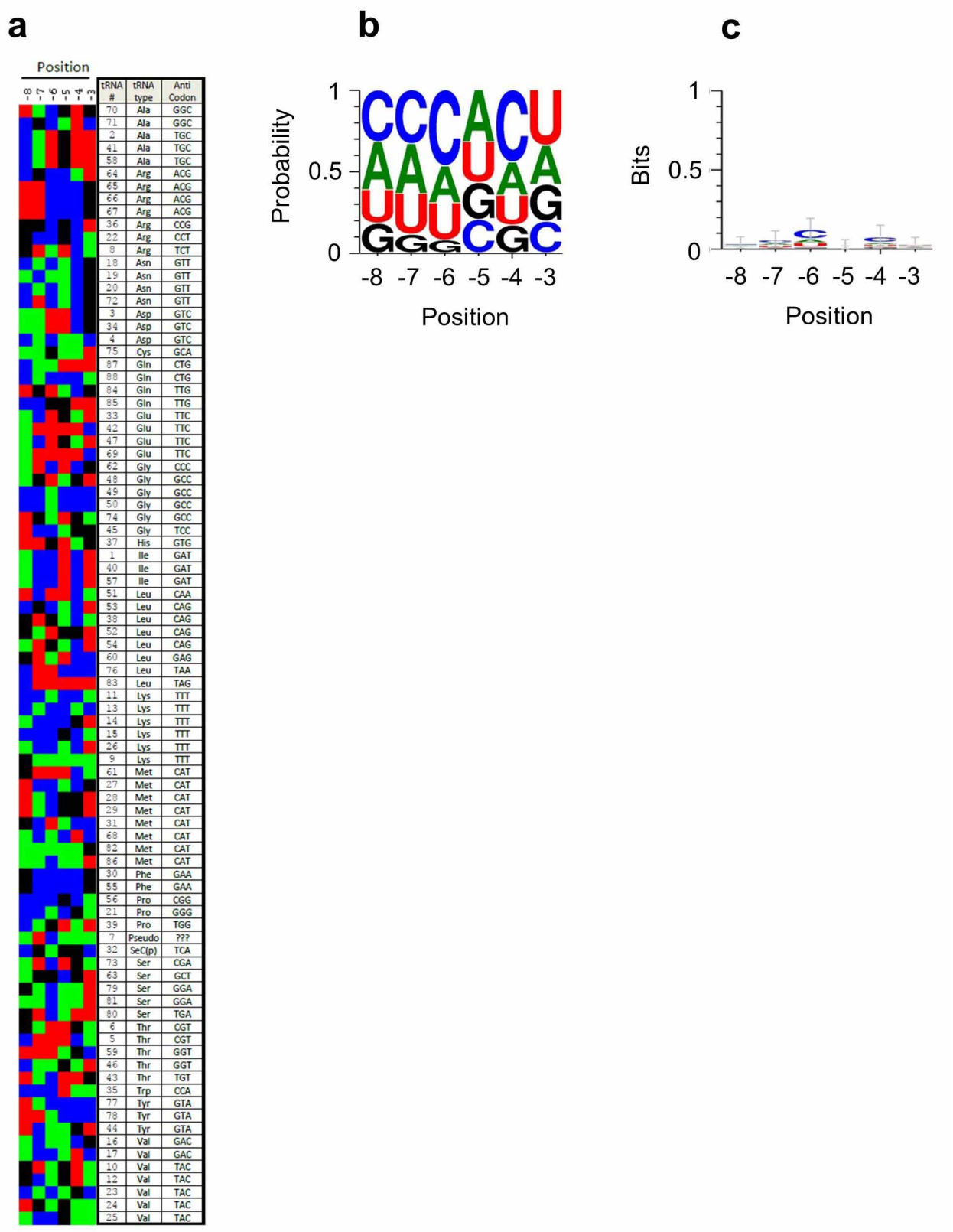
Position weight matrix. To examine how the position of different nucleobases in the binding site affects the reaction rate, we included the position of each base in the regression model. This model has 18 explanatory variables, three for each of the six positions, as explained above. Each variable is 1 if the respective base is at a given position, and 0 otherwise. For example, G4 will be 1 if the fourth base is a G, and 0 otherwise. We used the reference sequence as baseline, and did therefore not include these variables in the regression. Calculations of linear coefficients (Extended Data Fig. 7b), and the comparison between predicted and measured values for individual $\ln(k^{\mathrm{rel}})$ values (Fig. 3a) were performed as described above.

Functional couplings between bases. We created interaction variables of the same form as in the two-dimensional model 2, but these interaction terms were composed of two bases. For example, A1G4 is 1 if the first base is A and the fourth base is G, and will be 0 otherwise. We then performed calculations with the two-dimensional model 240 times, each time adding a separate interaction term. Interactions whose effect was statistically significant ($P < 10^{-5}$) were retained, other interactions were not considered further. Next, we included all statistically significant interactions in one model, which was further pared using stepwise regression[38]. This approach yielded a model similar to the position weight matrix augmented with the 44 most significant interaction terms. Calculations of linear coefficients (Fig. 3b), and the comparison between predicted and measured values for individual $\ln(k^{\mathrm{rel}})$ values (Fig. 3a) were performed as described above.

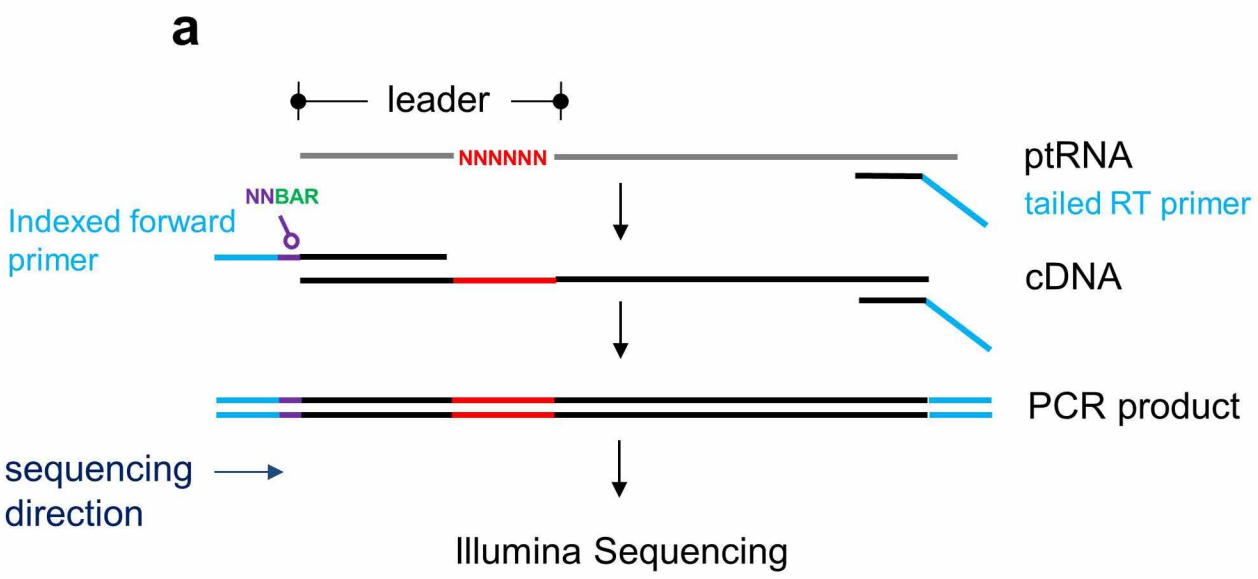
Neural network analysis. Analysis was performed with the MATLAB Neural Networks Toolbox (v. 3.0). Data input was identical to the two-dimensional model above. Data were fit to a three-layer feed-forward network with 13 hidden nodes.

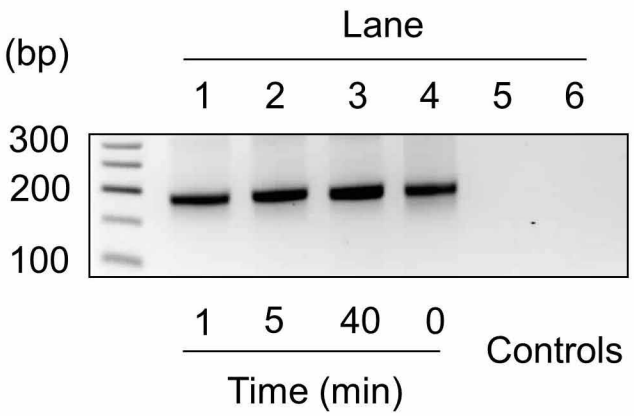
Interaction terms are not necessary in this model, because the neural network learns the interaction patterns from the raw sequence data. The resulting network was used to generate estimates for the reaction rate for each base sequence. Neural nets were trained on 60% of the data, validated on 20% of the data and tested on the remaining 20%. Almost identical R^2 values were obtained for both the 20% hold-out sample, and the entire data set.

31. Guo, X. *et al.* RNA-dependent folding and stabilization of C5 protein during assembly of the *E. coli* RNase P holoenzyme. *J. Mol. Biol.* **360,** 190–203 (2006).
32. Christian, E. L., McPheeters, D. S. & Harris, M. E. Identification of individual nucleotides in the bacterial ribonuclease P ribozyme adjacent to the pre-tRNA cleavage site by short-range photo-cross-linking. *Biochemistry* **37,** 17618–17628 (1998).
33. Cha, S. Kinetics of enzyme reactions with competing alternative substrates. *Mol. Pharmacol.* **4,** 621–629 (1968).
34. Chan, P. P. & Lowe, T. M. GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Res.* **37,** D93–D97 (2009).
35. Northrop, D. B. Fitting enzyme-kinetic data to *V/K. Anal. Biochem.* **132,** 457–461 (1983).
36. Northrop, D. B. Rethinking fundamentals of enzyme action. *Adv. Enzymol.* **73,** 25–55 (1999).
37. Theil, H. *Economic Forecasts and Policy* (North Holland Publishing, 1961).
38. Bendel, R. B. & Afifi, A. A. Comparison of stopping rules in forward ''stepwise'' regression. *J. Am. Stat. Assoc.* **72,** 46–53 (1977).



Extended Data Figure 1 | C5 binding site in the 87 ptRNA leaders in E. coli. **a**–**c**, Alignment and sequence logos for the C5 binding site in all 87 ptRNA leaders encoded by *E. coli*. Binding of C5 to the consecutive ptRNA positions −3 to −8 is well established, based on a crystal structure[9] and biochemical evidence[10]; that is, looping of bases seen for certain RNA- and DNA-binding proteins, does not occur with C5. Consistent with this idea, we did not detect any sequence motif with the MEME software, when including positions −1 to −10. **a**, Sequence alignment. Sequences were aligned with CLUSTAL. Coloured squares indicate the bases (C, blue; A, green; U, red; G, black). Anticodon, the anticodon recognized by the tRNA; tRNA#, the tRNA identification number; tRNA type, the amino acid. **b**, Sequence logo depicting the probability of any base at a given position, based on the alignment in **a**. The logo was generated with Weblogo. **c**, Sequence logo for the information content of the alignment in **a**. The logo was generated with Weblogo.

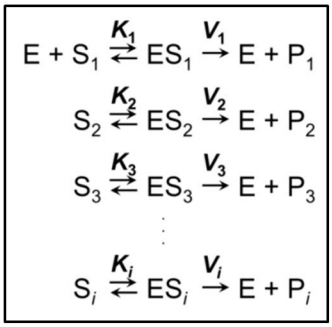




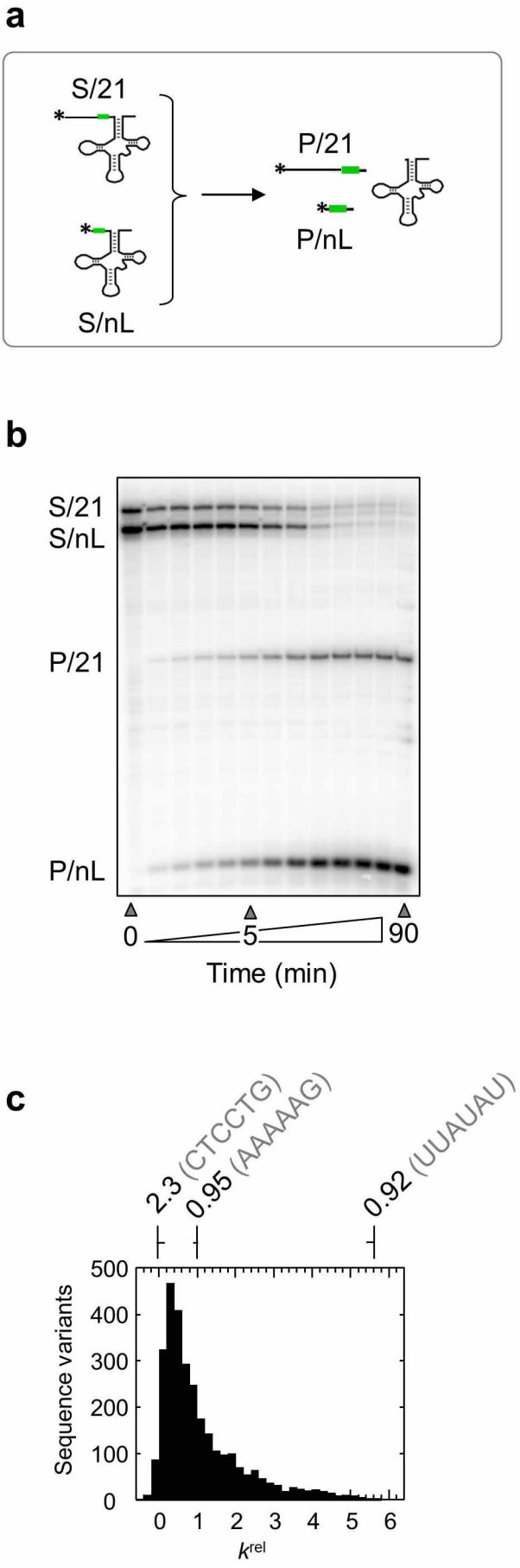
c

```
nt# 1 3 5                                  30   35
    | | |                                   |    |
5'  NNBARGGGAGACCGGAATTCAGATTGATGNNNNNNATGG 3'
```

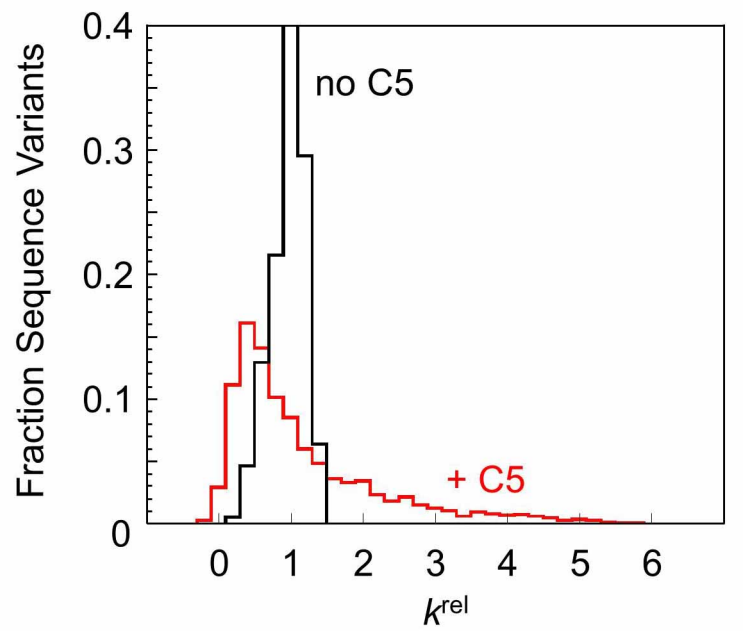
Extended Data Figure 2 | Preparation of DNA libraries for Illumina sequencing. **a**, BAR, the indexing barcode; NN, the degenerated barcode. For primer sequences see Methods. RT, reverse transcription. **b**, DNA libraries (PCR products, **a**) for samples at the time points indicated. Controls: lane 5, no RNA; lane 6, no reverse transcriptase. **c**, Read structure. Nucleotides 1 and 2 are degenerated barcode; nucleotides 3-5 are sample barcode (index tag); nucleotides 6–29 are additional leader sequence, nucleotides 30–35 are randomized leader sequence; nucleotides 38 onwards are tRNA.



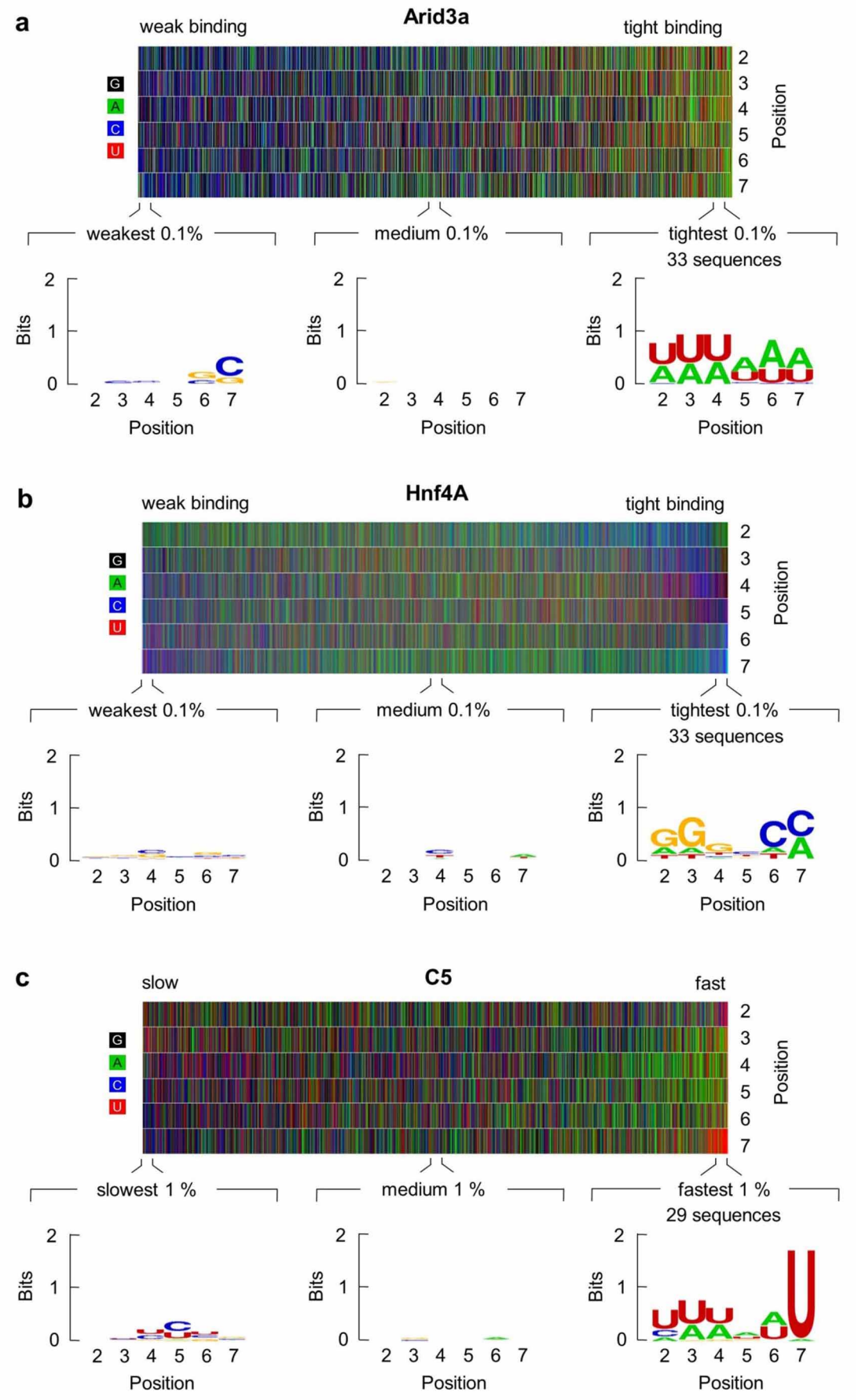
Extended Data Figure 3 | Multiple turnover reaction scheme. E, enzyme; $ES_{1...i}$, individual enzyme substrate complexes; $K_{1...i}$, individual functional binding constants; $S_{1...i}$, individual substrate variants; $V_{1...i}$, individual reaction rate constants.



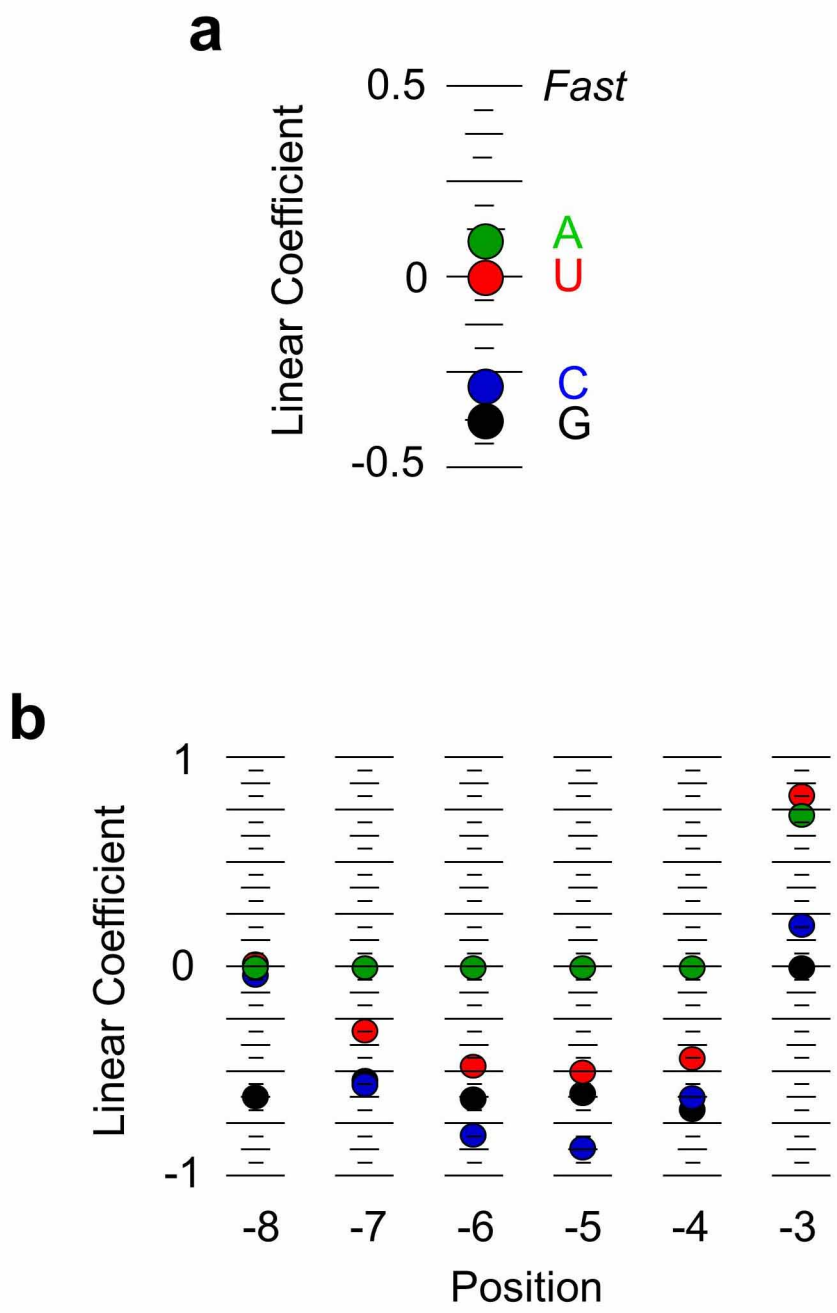
Extended Data Figure 4 | Effect of the 21 nucleotide extension on ptRNA processing by RNase P. **a**, Relative processing rate constants were measured for three sequence variants from different parts of the affinity distribution by PAGE. Reactions for each sequence variant were conducted in the presence of the randomized population (unlabelled) with equal amounts of substrate with (S/21) and without the 21-nucleotide extension (S/nL). The asterisk marks the position of the radiolabel at the 5′ end of the substrate. Reactions were conducted under the conditions described in the Methods. **b**, PAGE for the reaction of the reference sequence variant. The time point at 5 min is marked for reference. **c**, The effects of the 21-nucleotide extension on relative processing rate constants of the three indicated sequence variants. The position of each sequence variant in the affinity distribution of all sequence variants (Fig. 2d) is given for reference by the vertical line above the plot. The number indicates the factor (S/nL)/(S/21) by which the 21-nucleotide extension decreases the relative rate constant of the given sequence variant, given as average from three independent experiments. The horizontal line approximates the degree of the relative change. The 21-nucleotide extension decreases the observed for sequence variant (CTCCTG) by a factor of 2.3. For the genomically encoded leader sequence AAAAAG, the 21-nucleotide extension decreases k^{rel} for by a factor of 0.95; that is, the substrate with the extension reacts slightly faster than the substrate without extension. The fast reacting substrate (TTATAT) is also only minimally affected by the extension (0.92). Together, the data show only minor effects of the 21-nucleotide extension on the position of a given sequence variant in the affinity distribution.



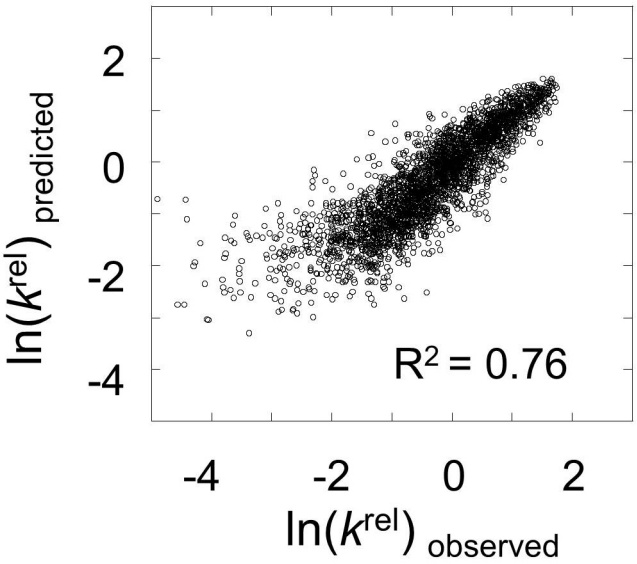
Extended Data Figure 5 | Processing of ptRNA$^{Met(-3-8)}$ by RNase P without C5. Distribution of k^{rel} values for processing of ptRNA$^{Met(-3-8)}$ by RNase P without C5 (black line). Data were obtained analogously to those with C5. For comparison, the distribution of k^{rel} values with C5 is shown (red line).



Extended Data Figure 6 | Sequence logos are only associated with the high-affinity tail of the distribution. **a**, Plot of sequence variants ranked from weakest to tightest binder to the specific transcription factor Arid3a (Fig. 2d), based on data published previously[18]. To facilitate direct comparison to the six-nucleotide binding site of C5, only approximately half of all sequences are shown in the plot, and only six positions (positions two to seven, as indicated) of the eight-nucleotide binding site are shown. The position in the binding site is marked on the right. The brackets mark 0.1% of sequence variants (33 sequences) that bind tightest, fall into the medium, and bind weakest. Sequence logos show the information content in these sequences. The logos were generated with Weblogo. Sequence signatures of the tightest binding variants are highly enriched in physiological substrates of Arid3a[18]. **b**, Plot of sequence variants ranked from weakest to tightest binder to another specific transcription factor, Hnf4a, based on data published previously[18]. Approximately half of all sequences are shown in the plot, and six positions (positions two to seven, as indicated) of the eight-nucleotide binding site. Sequence signatures of the tightest binding variants are highly enriched in physiological substrates of Hnf4A[18]. **c**, Plot of sequence variants ranked from slowest to fastest reacting for C5 (Fig. 2e). The brackets mark 1% of sequence variants that react fastest, fall into the medium and react slowest. Sequence logos were generated as in **a**.



Extended Data Figure 7 | Sequence determinants for substrate recognition by C5. **a**, Model considering identity, but not position of a given base in the C5 binding site. Ranking of the four bases according to their potential to promote (positive linear coefficient) or decrease (negative linear coefficient) functional C5 binding. For calculation of linear coefficients, see the Methods. **b**, Position weight matrix (PWM) model considering both base identity and position in the binding site, but assuming independent contributions of each position. The plot shows the ranking of the bases according to their potential to promote (positive linear coefficient) or decrease (negative linear coefficient) functional C5 binding, relative to the reference sequence (AAAAAG, Fig. 1c). Bases are coloured as in **a**. For the calculation of linear coefficients, see the Methods.



Extended Data Figure 8 | Neural network analysis. Correlation between observed k^{rel} and values calculated with the best model obtained by neural network analysis (Methods).

Extended Data Table 1 | Sequencing data.

Replicate 1				
Timepoint	0	T 1	T 5	T 40
(min:sec)	0:00	1:19	5:28	40:00
Fraction ptRNA processed *	0	0.14	0.28	0.55
Total sequence reads (number)	2,828,358	4,212,150	4,603,710	4,882,095
Reads passed quality threshold †	2,646,624	3,849,190	4,391,105	4,676,773
Fract. reads below quality threshold	0.064	0.086	0.0462	0.042

Replicate 2				
Timepoint	0	T 0.5	T 1	T 5
(min:sec)	0:00	0:37	1:15	5:15
Fraction ptRNA processed *	0	0.08	0.12	0.28
Total sequence reads (number)	6,434,248	8,172,493	11,054,769	11,604,173
Reads passed quality threshold †	5,800,933	7,476,616	10,341,132	10,421,304
Fract. reads below quality threshold	0.098	0.085	0.064	0.101

The overall number of reads obtained for the respective time points in the independent replicate experiments 1 and 2, and the number of reads that passed the quality control. *The fraction of processed ptRNA was determined from PAGE data (Fig. 1e and Methods). †Read numbers after filtering for potential PCR artefacts for each sequence variant (Methods).