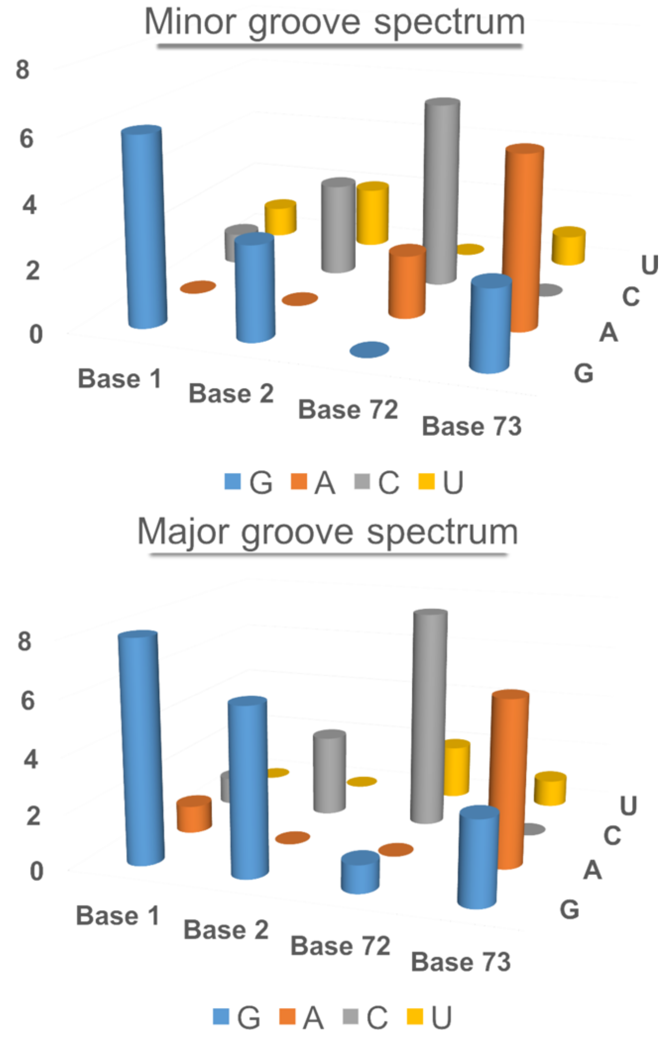
Supplement.

A. A consensus pattern in the topmost four tRNA acceptor-stem bases is almost the same for tRNAs recognized via each groove.



Comparison of the first four unique bases of the tRNA acceptor stems of the 20 isoaccepting tRNAs reveals why the basis of groove recognition presents so challenging a problem (Figure S1). The spectrum of bases for those tRNAs recognized by their minor groove is almost identical to that for tRNAs recognized via the major groove (this was previously noted (1)). In fact, the correlation between the spectrum of bases used in each of the four positions has an R^2 value = 0.79. This value is surprisingly high, given that the discrimination between aaRS Classes is so fundamental. One important implication is that it is very unlikely that the bases themselves—i.e. their main effects in the design matrix in Table S1—can explain the respective groove recognition properties without important higher-order interactions. Another is that the topmost four bases have a consensus composition {G, G/C, C, A}—i.e. nearly all tRNA acceptor stems have a G-C as their first base pair and A as Discriminator base, irrespective of which groove is recognized. This consensus composition suggests the ancestral minihelix recognized by both Classes began with a G-C base pair and reinforces the observation that the acceptor stem code for groove recognition is hidden by higher-order interactions.

Figure S1. Base composition spectra of the topmost four bases of the acceptor stems of tRNAs recognized by their minor and major grooves. The two histograms almost superimpose, and have an R^2 value of 0.79.

Table S1 Design matrix for regression analysis of groove recognition.

tRNA	Groove	β-branched	(-1)(Y/R)	(-1)G/A	1(Y/R)	1(G/A)	2(Y/R)	2(G/A)	72(Y/R)	72(G/A)	73(Y/R)	73(G/A)
Ala	**-1**	0	0	0	-1	1	-1	1	1	1	-1	-1
Cys	**1**	0	0	0	-1	1	-1	1	1	1	1	-1
Asp	**-1**	0	0	0	-1	1	-1	1	1	1	-1	1
Glu	**1**	0	0	0	-1	1	1	-1	1	1	-1	1
Phe	**1**	0	0	0	-1	1	1	1	1	1	-1	-1
Gly	**-1**	0	0	0	-1	1	1	1	1	1	1	-1
His	**-1**	0	-1	1	-1	1	1	1	1	1	-1	-1
Ile	**1**	1	0	0	0	0	-1	1	0	0	-1	-1
Lys	**-1**	0	0	0	-1	1	0	0	1	1	-1	-1
Leu	**1**	0	0	0	-1	1	1	1	1	1	-1	-1
Met	**1**	0	0	0	1	1	-1	1	-1	-1	-1	-1
Asn	**-1**	0	0	0	-1	-1	0	0	1	-1	0	0
Pro	**-1**	0	0	0	1	1	-1	1	-1	1	-1	-1
Gln	**1**	0	0	0	1	-1	-1	1	-1	-1	-1	1
Arg	**1**	0	0	0	-1	1	1	1	1	1	-1	0
Ser	**-1**	0	0	0	-1	1	-1	1	1	1	-1	1
Thr	**-1**	1	0	0	-1	1	1	1	1	1	-1	-1
Val	**1**	1	0	0	-1	1	0	0	1	1	-1	-1
Trp	**-1**	0	0	0	-1	-1	-1	1	1	-1	-1	1
Tyr	**-1**	0	0	0	-1	1	-1	1	1	1	-1	-1

Table S2. Computation of G_{calc}. Each cell underneath an amino acid shows the product of the estimate in the second column with the appropriate element of the design matrix in Table S1; totals are accumulated in the bottom row.

Class I

Term	Estimate	Cys	Glu	Ile	Leu	Met	Gln	Arg	Val	Trp	Tyr
Intercept	**-0.5**	-0.500	-0.500	-0.500	-0.500	-0.500	-0.500	-0.500	-0.500	-0.500	-0.500
β-branched	**-2.5**	0	0	-2.500	0	0	0	0	-2.500	0	0
(-1)(Y/R)	**2**	0	0	0	0	0	0	0	0	0	0
1(Y/R)	**0.5**	-0.500	-0.500	0	-0.500	0.500	0.500	-0.500	-0.500	-0.500	-0.500
2(Y/R)	**1**	-1.000	1.000	-1.000	1.000	-1.000	-1.000	1.000	0	-1.000	-1.000
2(G/A)	**1**	1.000	-1.000	1.000	1.000	1.000	1.000	1.000	0	1.000	1.000
72(G/A)	**-0.5**	-0.500	-0.500	0.000	-0.500	0.500	0.500	-0.500	-0.500	0.500	-0.500
β-branched*2Y/R	**-4**	0	0	4.000	0	0	0	0	0	0	0
2(Y/R)*73(Y/R)	**-1**	1.000	1.000	-1.000	1.000	-1.000	-1.000	1	0	-1.000	-1.000
1(Y/R)*72(G/A)	**-0.5**	0.500	0.500	0	0.500	0.500	0.500	0.500	0.500	-0.500	0.500
2(Y/R)*2(G/A)	**-1**	1.000	1.000	1.000	-1.000	1.000	1.000	-1.000	0	1.000	1.000
β-branched*72GA	**4.5**	0	0	0	0	0	0	0	4.5	0	0
Total		**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**-1.000**	**-1.000**

Class II

Term	Estimate	Ala	Asp	Phe	Gly	His	Lys	Asn	Pro	Ser	Thr
Intercept	**-0.5**	-0.500	-0.500	-0.500	-0.500	-0.500	-0.500	-0.500	-0.500	-0.500	-0.500
β-branched	**-2.5**	0	0	0	0	0	0	0	0	0	-2.500
(-1)(Y/R)	**2**	0	0	0	0	-2.000	0	0	0	0	0
1(Y/R)	**0.5**	-0.500	-0.500	-0.500	-0.500	-0.500	-0.500	-0.500	0.500	-0.500	-0.500
2(Y/R)	**1**	-1.000	-1.000	1.000	1.000	1.000	0	0	-1.000	-1.000	1.000
2(G/A)	**1**	1.000	1.000	1.000	1.000	1.000	0	0	1.000	1.000	1.000
72(G/A)	**-0.5**	-0.500	-0.500	-0.500	-0.500	-0.500	-0.500	0.500	-0.500	-0.500	-0.500
β-branched*2Y/R	**-4**	0	0	0	0	0	0	0	0	0	-4.000
2(Y/R)*73(Y/R)	**-1**	-1.000	-1.000	1.000	-1.000	1.000	0	0	-1.000	-1.000	1.000
1(Y/R)*72(G/A)	**-0.5**	0.500	0.500	0.500	0.500	0.500	0.500	-0.500	-0.500	0.500	0.500
2(Y/R)*2(G/A)	**-1**	1.000	1.000	-1.000	-1.000	-1.000	0	0	1.000	1.000	-1.000
β-branched*72GA	**4.5**	0	0	0	0	0	0	0	0	0	4.5
Total		**-1.000**	**-1.000**	**1.000**	**-1.000**	**-1.000**	**-1.000**	**-1.000**	**-1.000**	**-1.000**	**-1.000**

B. A major subset of amino acids obeys an identical model.

The twenty amino acids can be partitioned into a set of 16 that obey a subset of model in Table 2B with a complement of 4 that require additional coefficients. The four outliers are histidine for which the dominant discriminating feature is a G in the 5’ position, -1, plus the three amino acids with β-branched side chains. Thus, the set including Class I Leu, Met, Cys, Arg, Glu, Gln, Tyr, and Trp and Class II Ala, Asn, Asp, Gly, Lys, Ser, Pro, and Phe all obey a subset of the model in Table 2B with only seven coefficients plus the intercept. That model therefore has 8 degrees of freedom, the same number as the model in Table 2B. Moreover, the coefficients are identical for the two models, and the subset model gives R^2 = 1.0.

C. Simpler models fit a large subset and suggest a mechanistic foundation.

An even simpler model for the major partition has only two coefficients: 1(Y/R) and 2(Y/R)*73(Y/R), yet predicts the groove recognized by the reduced tRNA set with R^2 = 0.73. Coefficients for the two predictors {0.73, -0.98} are tolerably close to those of the full model {0.5, -1.0}. This reduced model depends only on the placement of pyrimidines and purines within the topmost four bases, and may furnish a basis for improved understanding of the mechanism by which these bases dictate which groove is recognized.

Removing outliers (Lys, Asn, Pro) from within that set of 16 leaves a set of 13 amino acids for which the same two predictors fit perfectly (R^2 = 1.0; logWorth ~ 80) with digital coefficients, {1, 1, -1}. This model uses only three degrees of freedom to fit 13 tRNA isoacceptors, and thus has 10 degrees of freedom. Curiously, the three amino acids with aromatic side chains remain within the set predicted by this very simple model. The thirteen qualifying tRNA isoacceptors thus contain representatives from each of the three recognized subclasses in each Class (Table S2). The intercept of this model is 1.0, which implies that minor groove recognition is the default for the subset. Computation of groove recognition is given in Table S3, which is similar in intent to Table S2.

These simple rules (Table S4) appear to underlie the more complex set of rules in Table 4. Moreover, increasing the set of correct predictions requires three, orthogonal sets of coefficients. Including His requires simply -1,Y/R; including the β-branched side chains requires {β-branched; β-branched*2,Y/R; and β-branched*72,≡/=}; and including Asn, Lys, Pro requires {72,≡/=; 1,Y/R*72,≡/=; 2,≡/=; 2,Y/R ; and 2,Y/R*2,≡/=). The values of the coefficients 4 in these marginal models are more or less identical to those from the full model in Table, suggesting that there is no interaction between them when combined. Thus, they represent independent accretions to the operational RNA code, as suggested in Figure 5. Moreover, the fact that all sets of coefficients have integral or half-integral values reinforces the conclusion that this is a digital code.

Table S3. Groove computation with 3 coefficients for a reduced amino acid set. This table is identical in intent to Table S2. Here, however, the bases themselves are also tabulated, together with the products of their coefficients and the corresponding element of the design matrix. Inspection of Figure 1 will confirm the entries in columns 3-7. The intercept for this model is 1, the coefficients are: -1(Y/R) = 2, 1(Y/R) = 1, and 2(Y/R)*73(Y/R) = -1, leading to the respective contributions in columns 9-11 to the total in column 12.

Class I			**Base 1**	**Base 2**	**Base 72**	**Base 73**	**Constant**	**-1**	**1Y/R**	**2Y/R*73Y/R**	**Total**
IA	**Arg**		G	C	C	G	1	0	-1	1	**1**
IA	**Cys**		G	G	C	U	1	0	-1	1	**1**
IA	**Leu**		G	U	C	G	1	0	-1	1	**1**
IA	**Met**		U	G	A	G	1	0	1	-1	**1**
IB	**Glu**		G	C	C	A	1	0	-1	1	**1**
IB	**Gln**		C	G	A	A	1	0	1	-1	**1**
IC	**Tyr**		G	G	C	A	1	0	-1	-1	**-1**
IC	**Trp**		A	G	U	G	1	0	-1	-1	**-1**
Class II											
IIA	**Ala**		G	G	C	A	1	0	-1	-1	**-1**
IIA	**Gly**		G	G	C	G	1	0	-1	-1	**-1**
IIA	**His**	**G**	G	C	C	A	1	-2	-1	1	**-1**
IIA	**Ser**		G	C	C	U	1	0	-1	-1	**-1**
IIB	**Asp**		G	G	C	G	1	0	-1	-1	**-1**
IIC	**Phe**		G	C	C	A	1	0	-1	1	**1**

D. Thermodynamic relationships associated with stacking of bases 1, 2, 72, and 73.

The rules arising from the two coefficients of the minimal model for groove recognition (Table S4) suggest their origin in RNA base stacking energetics. Base stacking free energies appear to be the most important determinants of helical stability at nicked or gapped DNA (2,3), and hence likely at the termini of double helical RNA stems.

We used two methods to assess a possible thermodynamic basis for the patterns in Table S4.

(i) We summed directly the estimates of base stacking as estimated by Frank-Kamenetskii (2,3). These values are potentially useful because they are provided for interactions between single bases, not base pairs. Thus, they enable an direct estimation of combinatorial energetic differences associated with the rules in Table S4. However, stacking energies for RNA differ significantly from those for DNA, especially for uridine and thymine (4). R^2 values for the correlations between computed RNA and DNA stacking energies for A, C, and G (0.89) and separately for T(U) (0.8) are both >> 0.54, the overall correlation between computational and experimental values. Thus, we scaled each set separately to generate a consistent set of RNA stacking energies. Estimates were derived by summing <C+U>- and <G+A>-averaged 5' – 3' stacking free energies of the five unique 3'-terminal bases of the acceptor stem (1,2,71,72,73) each observed acceptor-stem configuration and are compiled, together with that of the Discriminator base, in Table 5. The free energy difference, $\Delta(\Delta G)_{stck,}$ between extended and hairpin configurations for each combination consistent with rules in Table S4 was then estimated by subtracting the stacking energy of the Discriminator base itself from the total stacking free energy. The $\Delta(\Delta G)_{stck}$ values for tRNAs recognized via the minor groove were less stable by –1.2 kcal/mole, and the Student t-test for this difference had a P-value of 0.0009. Base-stacking configurations that form the Class I 3'-terminal CCA hairpin are, on average, significantly easier to break than those associated with recognition from the major groove.

(ii) We took each unique complete pattern of bases from Figure 4 as the topmost four bases in a minihelix in which the core bases were drawn from a polyG-polyC minihelix derived from $tRNA^{Leu}$. We then submitted each minihelix, together with the core to Mfold (5) and subtracted the folding energies of the core from the total energies to estimate the stability of the terminal seven bases. Values estimated using Mfold (5) (Table 5) were qualitatively similar, but their variance was too great to conclude statistical significance.

Table S4 Structural biology data summary

Amino acid	Class	PDB ID	Organism	Chain	RMS, Å
Gln	**1**	**2RE8**	***E coli***	**B**	**0.00**
Ile	**1**	**1QU3**	***St. aureus***		
Cys	**1**	**1U0B**	***E coli***	**A**	**0.693**
Leu	**1**	**5OMW**	***E coli***	**B**	**0.487**
Glu	**1**	**3AKZ**	***T maritima***	**E**	**0.575**
Arg	**1**	**1F7U**	***C cerevisiae***	**B**	**0.476**
Trp	**1**	**2DR2**	***H sapiens***	**B**	**0.527**
Ala	**2**	**3WQY**	***A fulgidis***	**C**	**0.872**
Asp	**2**	**1C0A**	***E coli***	**B**	**0.529**
Gly	**2**	**5E6M**	***H sapiens***	**C**	**0.733**
Thr	**2**	**1QF6**	***E coli***	**B**	**0.876**

References

1. Limmer, S., Hofmann, H.-P., Ott, G. and Sprinzl, M. (1993) The 3'-terminal end (NCCA) of tRNA determines the structure and stability of the aminoacyl acceptor stem. *Proc. Natl. Acad. Sci. USA*, **90**, 6199-6202.

2. Yakovchuk, P., Protozanova, E. and Frank-Kamenetskii, M.D. (2006) Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. *Nucleic Acids Res.*, **34**, 564–574

3. Krueger, A., Protozanova, E. and Frank-Kamenetskii, M.D. (2006) Sequence-Dependent Basepair Opening in DNA Double Helix. *Biophys. J.*, **90**, 3091–3099.

4. Brown, R.F., Andrews, C.T. and Elcock, A.H. (2015) Stacking Free Energies of All DNA and RNA Nucleoside Pairs and Dinucleoside Monophosphates Computed Using Recently Revised AMBER Parameters and Compared with Experiment. *J. Chem. Theory Comput.* , **11**, 2315−2328.

5. Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research*, **31**, 3406–3415.