Supplementary information

Captions for supplementary figures

Figure S1: PKI structurally and functionally mimics a codon-anticodon helix
Comparison of the base pairing between the tRNA anticodon numbered 34, 35, 36 with a cognate codon numbered 1, 2, 3 shown on the left and the pseudoknot PKI from CrPV Intergenic IRES. Structure of the Anticodon Stem Loop or ASL (in orange) base paired to a cognate codon in the A site of the ribosome (on the left) and structure of PKI in the A site of the ribosome (on the right) are represented from respectively X-ray (pdb:5e81) (16) and CryoEM (pdb: 5it9) (33) molecular models. The tRNA anticodon part is in red and the mRNA codon part in cyan. In the IRES structure, the 5'-end of the "mRNA" continues to pair with the "anticodon 37 and 38". A library containing the 4096 codon-anticodon combinations in the frame of PKI was generated.

Figure S2: Ultrahigh-throughput screening of randomized IRES gene libraries.
(a) Schematic of the droplet-based microfluidic screening pipeline. The screening pipeline operates in three main steps. First, the genes of the library are diluted into a PCR mixture prior to being individualized into small picoliter-sized water-in-oil droplets (left device). The emulsion is collected off-chip into a conventional PCR tube and the whole emulsion is thermocycled into a regular thermocycler. Next, the amplified DNA-containing droplets (orange) are reinjected into a fusion device (middle device) and synchronized with on-chip generated droplets (red) containing an *in vitro* expression mixture (a Rabbit Reticulocyte Lysate supplemented in T7 RNA polymerase in this study). Both sets of droplets are synchronized, fused by an electric field and collected off-chip. Genes are then expressed during an off-chip incubation step prior to reinjecting the droplets into a last device (right device) aiming at analyzing droplets fluorescence and sorting them accordingly. **(b)** Occurrence frequency matrix of the 4096 variants contained in the starting gene library. The occurrence frequency has been calculated for each variant contained in the starting library and the value was color-coded accordingly. The ~ 6-fold occurrence difference between the most and the least represented variant is supportive of a generally unbiased the library and confirmed that a ~ 37-fold coverage was enough to see

each variant during the first rounds of screening. **(c)** Representation of the evenness of the variants contained in the starting library. Sequences were ordered and numbered according to their occurrence in the library (a sequence ID inversely proportional to the occurrence was attributed each sequence) prior to being plotted as a function of their cumulative occurrence. The straight alignment of the point is indicative of an overall even representation of the different variants. **(d)** GFP fluorescence profile of the droplets from the two replicates of the experiment. In each experiment, the dashed line indicates the lower limit of the sorting gate. The sorting gate was set such that every droplet with a fluorescence detached from the negative population was recovered. **(e)** Representation evenness of the variants selected in each replicate. Sequences were ordered and numbered according to their occurrence in the library (a sequence ID inversely proportional to the occurrence was attributed each sequence) prior to being plotted as a function of their cumulative occurrence. Sequences corresponding to molecules initially individualized and amplified several hundreds of times in droplets are expected to be over-represented in comparison to those coming from a rare mutation event during PCR or from a sequencing error. Therefore, whereas the formers are expected to accumulate at a high, the latter should accumulate more slowly and a biphasic curve like those observed here is expected. Consequently, the breakpoint at the junction of both curves corresponds to the threshold between the relevant sequences (signal) and the non-relevant ones (noise). The Venn diagram represents the combinations that are present in both replicates.

Figure S3: Combinations of codon-anticodon containing Watson-Crick base pairs that are selected

The combinations are ranked according the 3-34 pair and their proportion is shown in parentheses. Interestingly, the combinations Y3/R34 are two times more frequent than the combinations R3/Y34. The amino acids coded by the corresponding codons are shown under each combination. No combination was selected for Methionine, Glutamic acid, Lysine and the three stop codons.

Figure S4: Combinations of codon-anticodon containing Watson-Crick base pairs that are not selected

The combinations are ranked according the position of A/U base pairs (shown in blue). Start (green dot) and stop codons (red dots) are indicated. The proportion of each category is shown under the figure. The two pie charts represent the proportion of nucleotides at position 34 and 3 in these missing combinations.

Figure S5: Combinations of selected codon-anticodon containing Watson-Crick base pairs with one or two mismatches

(a) The codon-anticodon combinations are listed according to the position of the mismatches. There are 56 combinations with one mismatch and three with two. The mismatches are colour-coded (G/U in green, G/A in orange, C/A in yellow, C/U in brown, A-A in dark grey, U-U in purple and G-G in light grey). The proportion and position of each mismatch are summarized at the bottom. About 2/3 of mismatches occur at the third “wobble” position and the most frequent are G/U (12), G/A (10), C/A (6) and C/U (5).

(b) The histogram represents the number of mismatches obtained for each of the three positions of the codon. The pie charts represent the proportion of each of the four nucleotides at position 34 in the codon-anticodon combinations containing mismatches at position 1 and 2, and at position 3 of the codon.

(c) The histograms represent the total number of each type of mismatch at the three positions of the codon.

Figure S6: Combinations of selected codon-anticodon interacting with one or two mismatches.

(a) The histogram represents the number of G/U (green), C/A (yellow), G/A (orange) and U/C mismatches (brown) found in each of the 3 positions of the codon. Striped bars represent the orientations 1-36, 2-35 and 3-34 and full bars represent the opposite orientations. The pie charts represent the proportion of each orientation in the 3 positions of the codon.

(b) The histogram represents the relative Renilla luciferase activity obtained with PKI in frame of a Renilla luciferase reporter with several codon-anticodon combinations. Combinations that were selected through the microfluidic pipeline are indicated by (+) and combinations that were not selected are indicated by (-).

(c) Experimental validation of the preferred orientation of G/U mismatches at the position 3 of the codon. The histogram represents the relative Renilla luciferase activities of codon-anticodon combinations in PK1 with G/U mismatches in both orientation at the three positions of the codons and in frame with a Renilla coding sequence.

Figure S7: Second selection on the combinations in 'stringent' conditions.

(a) GFP fluorescence profile of the droplets from the two replicates of the selection in stringent conditions. In each experiment, the dashed line indicates the lower limit of the sorting gate. Representation evenness of the variants selected in each replicate is shown below the fluorescent profiles. Sequences were ordered and numbered according to their occurrence in the library (a sequence ID inversely proportional to the occurrence was attributed each sequence) prior to being plotted as a function of their cumulative occurrence. The straight alignment of the point is indicative of an overall even representation of the different variants. The Venn diagram represents the combinations that are present in both replicates. **(b)** The sequences of the active codon-anticodon pairs that are efficiently recognized by the ribosome in 'stringent' conditions are plotted on a matrix. The 64 codons are represented on the x-axis and 64 anticodons are represented on the y-axis. The nucleotides of the codons are numbered 1, 2 and 3 from 5' to 3'. The nucleotides of the anticodons are numbered 34, 35, 36 from 5' to 3' according to their position in tRNAs. The active codon-anticodon combinations are represented on the matrix by black squares (Watson-Crick pairs are along the diagonal) and by coloured squares for combinations containing mismatches that are parallel to the diagonal. The total number of hits is indicated on the upper right part of the matrix. The number of selected codon-anticodon pairs containing A, C, G and U at position 34 (anticodon) are shown on the right of the matrix. The number of selected codon-anticodon pairs containing A, C, G and U at position 3 (codon) are shown above the matrix. (**c**) Histogram representing the number of mismatches obtained in each of the three positions of the codons. The proportion of each of the four nucleotides at position 34 is shown in the pie charts for mismatches at positions 1 and 2 and at position 3.

Figure S8: The anticodons with a G34 are prone to miscoding

The 51 active codon-anticodon pairs containing G34 that are selected in the 'Relaxed' selection procedure are plotted on the matrix. Combinations containing

Watson-Crick base pairs are represented by black squares (15) and combinations containing mismatches (36) are highlighted using coloured squares according to the figure legend. The combinations that would lead to miscoding (and not G/U wobbling) are circled. The codons that are incorrectly decoded are indicated above the matrix and the anticodons that induce miscoding are shown on the right. The resulting miscoding events would, in a natural system, induce incorporations of non-cognate amino acids (in black) instead of the cognate amino acids (in red). An asterisk indicates the miscoding combination (Pro>Ser or $_{1}$CCC$_{3}$/$_{34}$GGA$_{36}$) that has been validated with a reporter gene by Mass Spectrometry (see S10).

Figure S9: Evolutionary clearance of G34 containing tRNAs in eukaryotic tRNAs of 3- and 4-codon boxes.

The heat map represents the ratio of the number of each of the putative tRNA gene corresponding to the 64 anticodons divided by the total number of tRNA genes in various eukaryotic genomes. The colour code is indicated at the bottom of the figure from black for abundant tRNA genes to white for tRNA genes that are absent from eukaryotic genomes. The species are indicated at the top of the figure and the anticodon for each tRNA gene, the corresponding codon and the amino acid identity are shown in the table on the right part of the figure. The orange boxes indicate that the tRNA genes containing anticodon starting with an A at position 34 are rare or absent. The yellow boxes indicate that the tRNA genes containing an anticodon starting with a G at position 34 are rare or absent. The cartoon on the right part summarizes the results on the heat map. In eukaryotes, the A34-containing tRNA genes were cleared throughout evolution in 2-box tRNA sets and the G34-containing tRNA genes have been cleared in 3- and 4-box tRNA sets (with the exception of Gly, a 4-box tRNA in which A34 has been cleared to favour G34). In eukaryotes, A34 is modified into I34; the clearance of A34 in 2-codon boxes results from the miscoding potential of I (that can pair with C, U, and A)(59).

Figure S10: Anticodons containing G34 do promote miscoding in Rabbit Reticulocyte Lysates.

The Renilla luciferase protein produced in rabbit reticulocyte lysate in the presence of *Homo sapiens* tRNA$^{Ala}_{GNN}$ transcripts and synthetic Renilla mRNA was purified via its C-terminal HA-tag, digested by trypsin and analysed by Mass Spectrometry. **(a)**

Mass Spectrometry analysis of the peptide sequence from Renilla luciferase protein produced in presence of *Homo sapiens* tRNA^{Ala}$_{GGA}$. The upper panel represents the wild-type peptide ($_{253}$MFIESDPGFFSNAIVEGAK$_{271}$) containing the expected Serine residue highlighted in yellow at the position of the Serine UCC codon inserted by endogenous tRNA^{Ser}. The lower panel shows the same peptide where the Serine was substituted with an Alanine residue (also in yellow) by the exogenous *Homo sapiens* tRNA^{Ala}$_{GGA}$ transcript confirming that this synthetic chimera is efficiently alanylated by endogenous AlaRS and picked up by ribosomes available in Rabbit Reticulocyte Lysates. **(b)** Mass Spectrometry analysis of the same reporter peptide. The upper panel shows the sequence of the peptide containing the expected Proline residue highlighted in yellow at the position of the Proline CCC codon inserted by endogenous tRNA^{Pro}. The lower panel shows the peptide where the Proline was substituted with Alanine (also in yellow) by the exogenous *Homo sapiens* tRNA^{Ala}$_{GGA}$ confirming that this anticodon supports miscoding when a C-A mismatch is present at the first position of the codon.

Figure S1

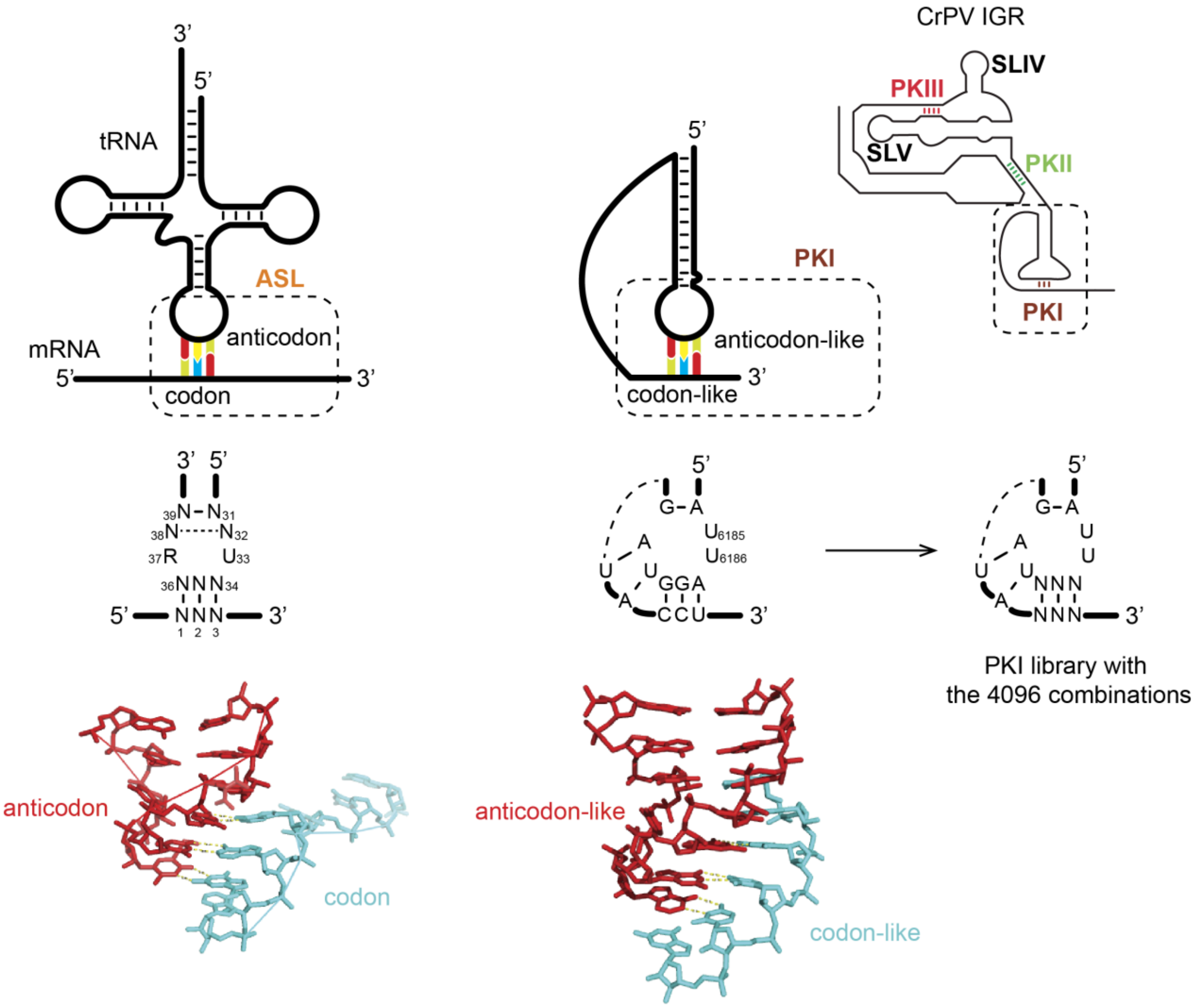

3’
5’
tRNA
ASL
anticodon
mRNA
5’
3’
codon
CrPV IGR
SLIV
PKIII
SLV
PKII
PKI
5’
PKI
anticodon-like
3’
codon-like
3’ 5’
39N - N31
38N······N32
37R U33
36N N N34
5’ N N N 3’
1 2 3
5’
G - A
U6185
A
U6186
U
U GGA
A
CCU 3’
5’
G - A
U
A
U
U
U NNN
A
NNN 3’
PKI library with
the 4096 combinations
anticodon
codon
anticodon-like
codon-like

Figure S2

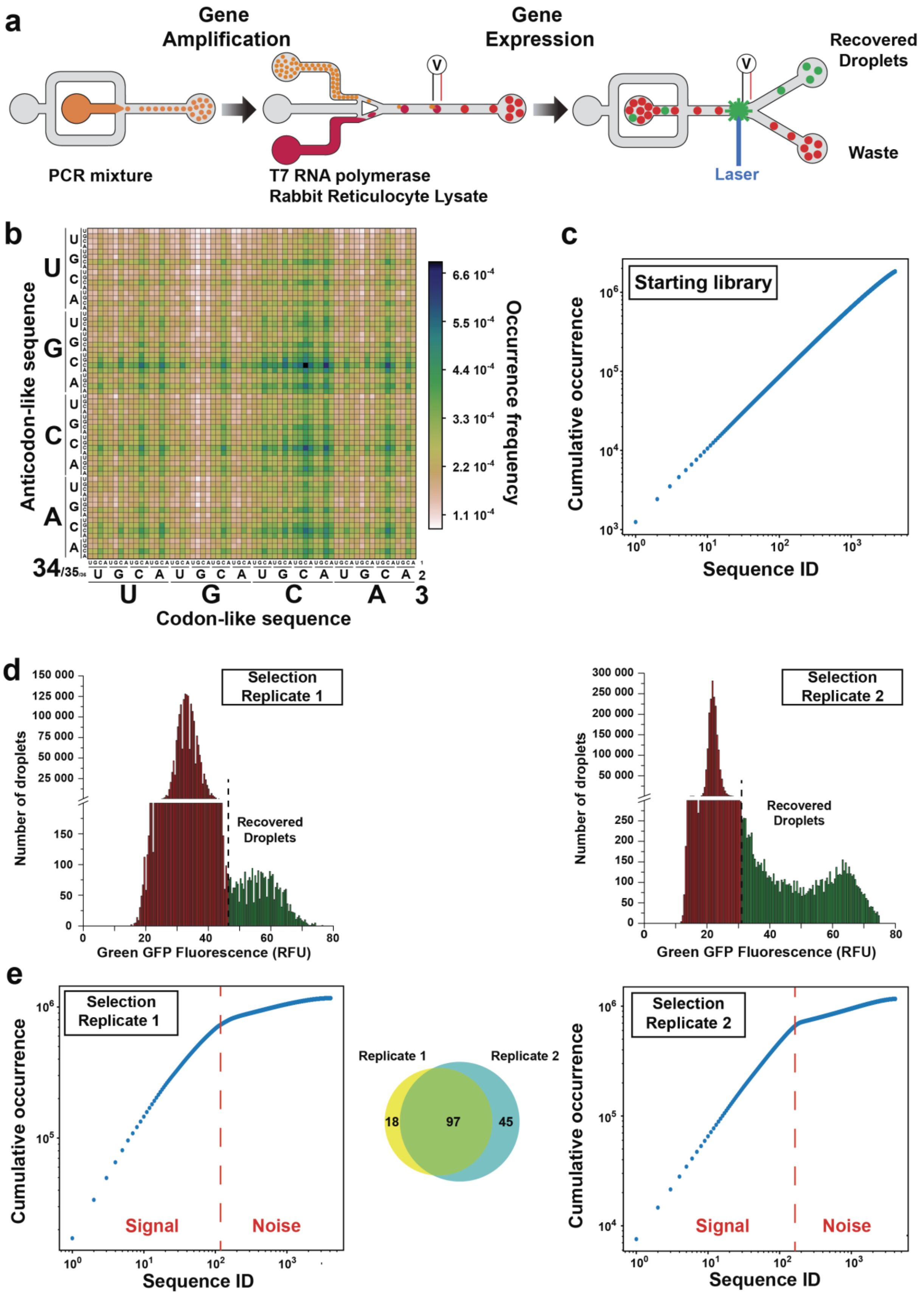

a
Gene Amplification
Gene Expression
Recovered Droplets
V
V
Waste
PCR mixture
T7 RNA polymerase
Rabbit Reticulocyte Lysate
Laser
b
Anticodon-like sequence
U
G
C
A
34/35/36
U G C A U G C A U G C A U G C A
U
G
C
A
Codon-like sequence
1
2
3
6.6 10⁻⁴
5.5 10⁻⁴
4.4 10⁻⁴
3.3 10⁻⁴
2.2 10⁻⁴
1.1 10⁻⁴
Occurrence frequency
c
Starting library
Cumulative occurrence
10³
10⁴
10⁵
10⁶
10⁰
10¹
10²
10³
Sequence ID
d
Selection Replicate 1
Number of droplets
150 000
125 000
100 000
75 000
50 000
25 000
150
100
50
0
0
20
40
60
80
Recovered Droplets
Green GFP Fluorescence (RFU)
Selection Replicate 2
Number of droplets
300 000
250 000
200 000
150 000
100 000
50 000
250
200
150
100
50
0
0
20
40
60
80
Recovered Droplets
Green GFP Fluorescence (RFU)
e
Selection Replicate 1
Cumulative occurrence
10⁶
10⁵
10⁰
10¹
10²
10³
Signal
Noise
Sequence ID
Replicate 1
Replicate 2
18
97
45
Selection Replicate 2
Cumulative occurrence
10⁶
10⁵
10⁴
10⁰
10¹
10²
10³
Signal
Noise
Sequence ID

Figure S3

38 Combinations with Watson-Crick base pairs

missing combinations for : Met, Glu, Lys, stop codons

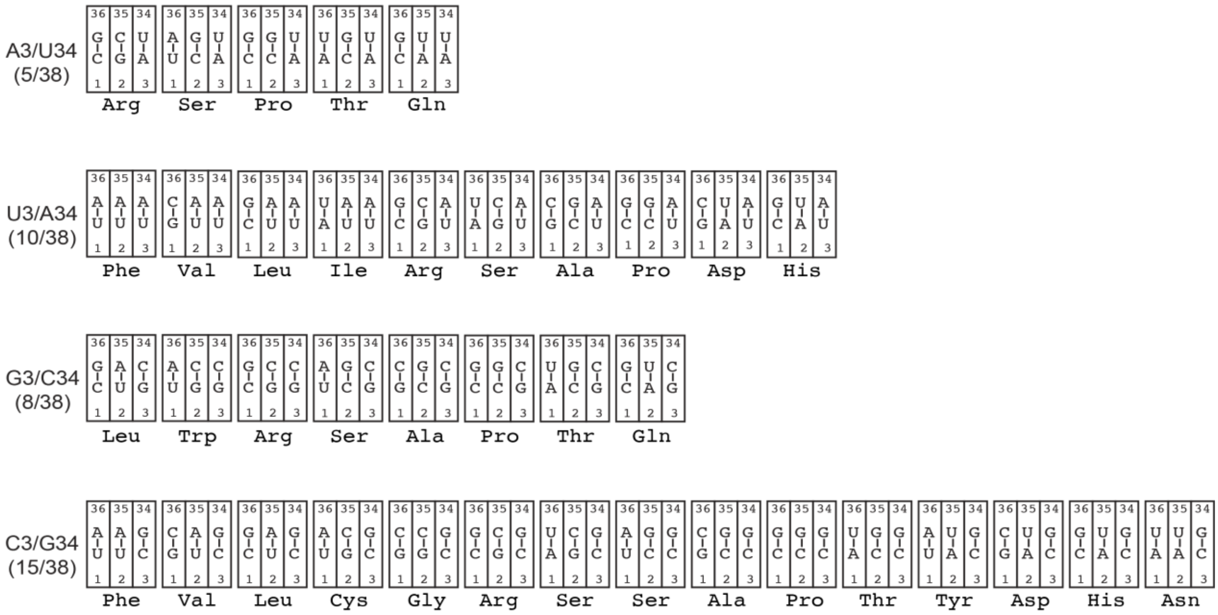


Figure S4

26 missing combinations with Watson-Crick base pairs

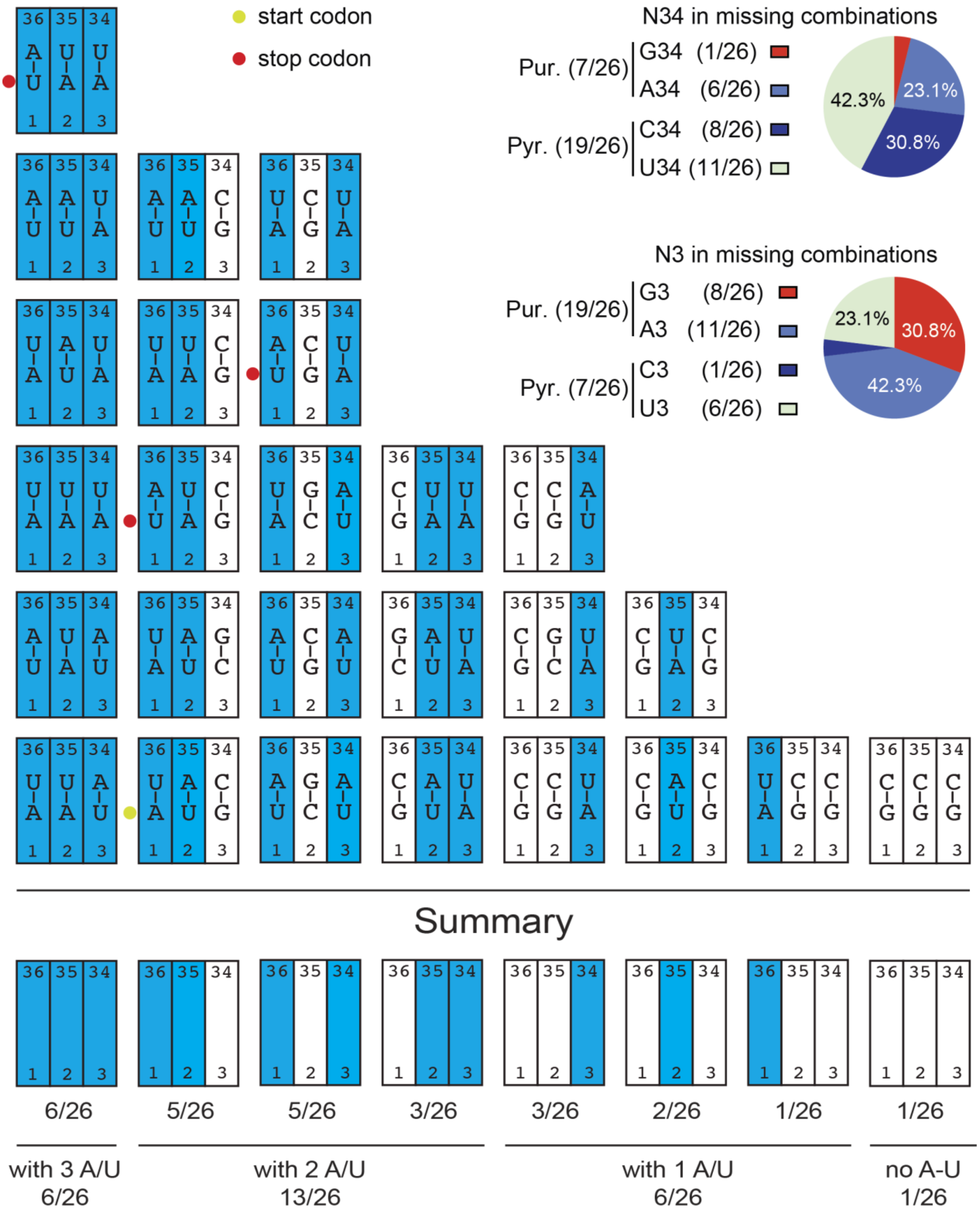


Figure S5a

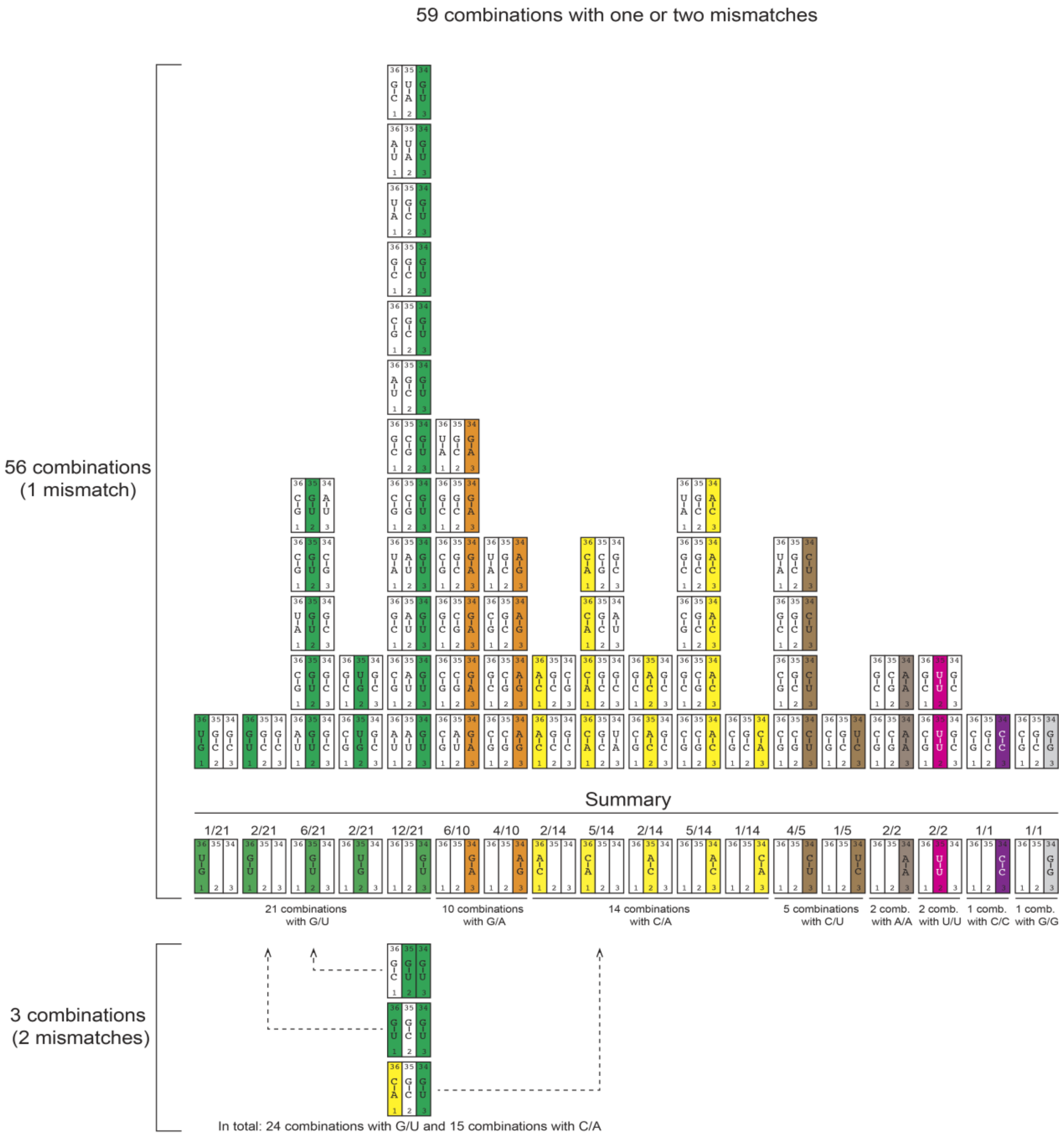

59 combinations with one or two mismatches
56 combinations
(1 mismatch)
3 combinations
(2 mismatches)
Summary
1/21
2/21
6/21
2/21
12/21
6/10
4/10
2/14
5/14
2/14
5/14
1/14
4/5
1/5
2/2
2/2
1/1
1/1
21 combinations
with G/U
10 combinations
with G/A
14 combinations
with C/A
5 combinations
with C/U
2 comb.
with A/A
2 comb.
with U/U
1 comb.
with C/C
1 comb.
with G/G
In total: 24 combinations with G/U and 15 combinations with C/A

Figure S5

Combinations with mismatches

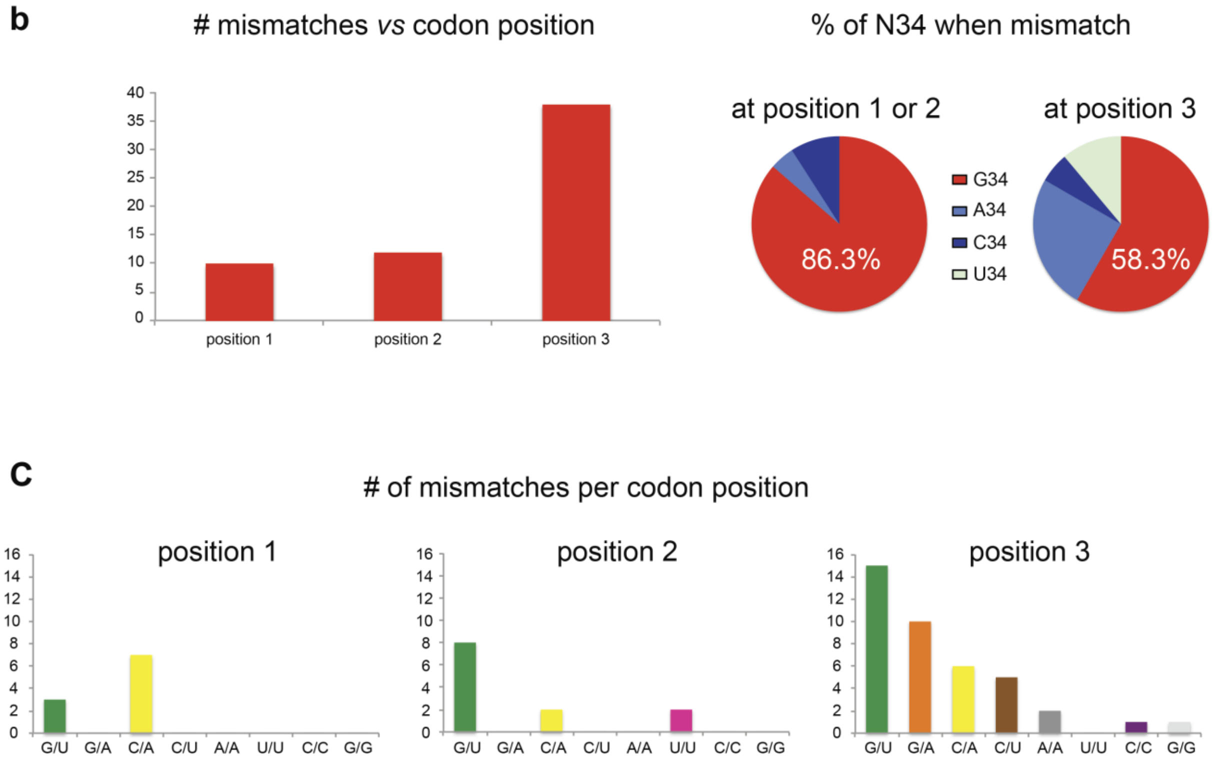


Figure S6a

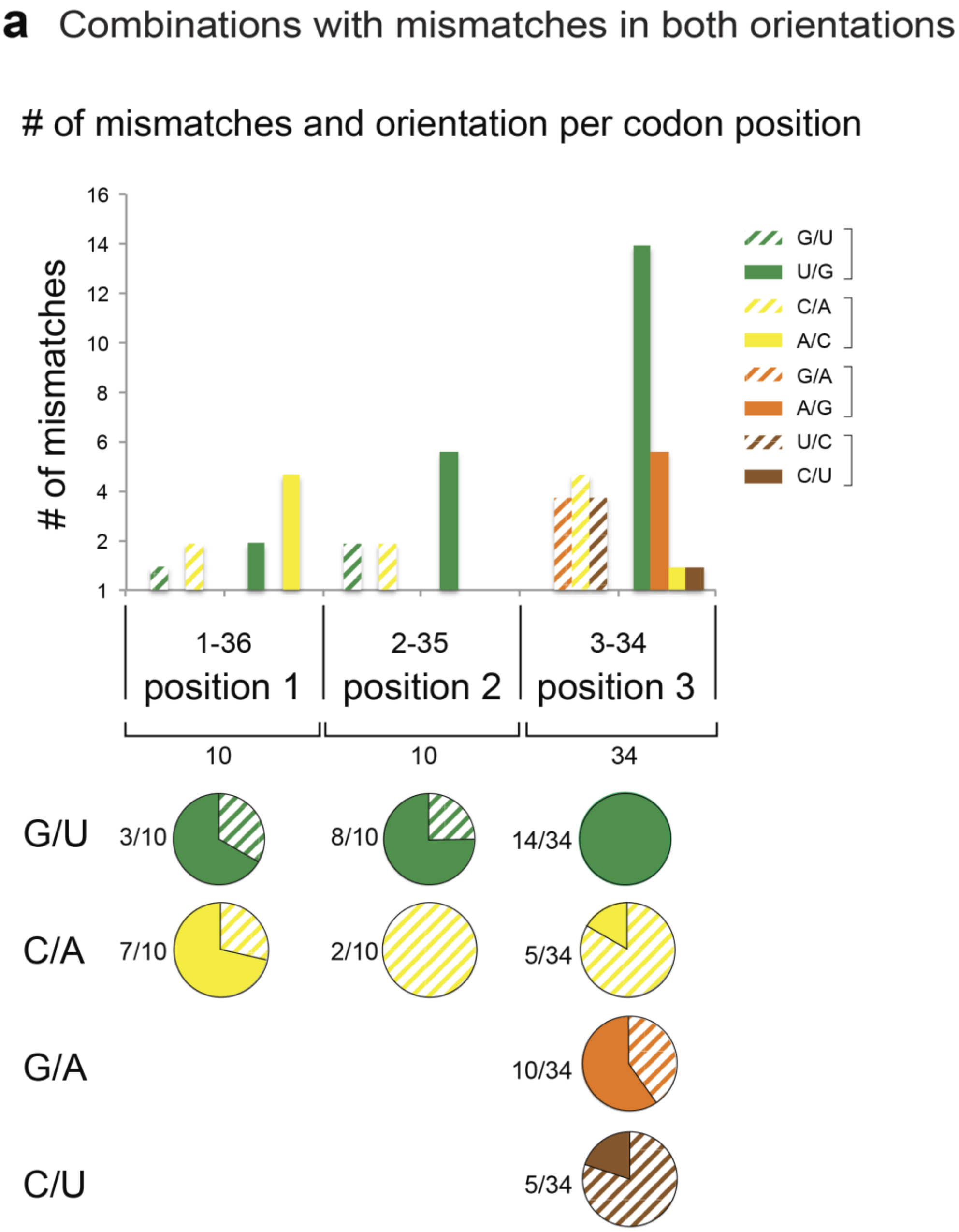

a
Combinations with mismatches in both orientations
# of mismatches and orientation per codon position
# of mismatches
16
14
12
10
8
6
4
2
1
G/U
U/G
C/A
A/C
G/A
A/G
U/C
C/U
1-36
position 1
2-35
position 2
3-34
position 3
10
10
34
G/U
3/10
8/10
14/34
C/A
7/10
2/10
5/34
G/A
10/34
C/U
5/34

Figure S6b-c

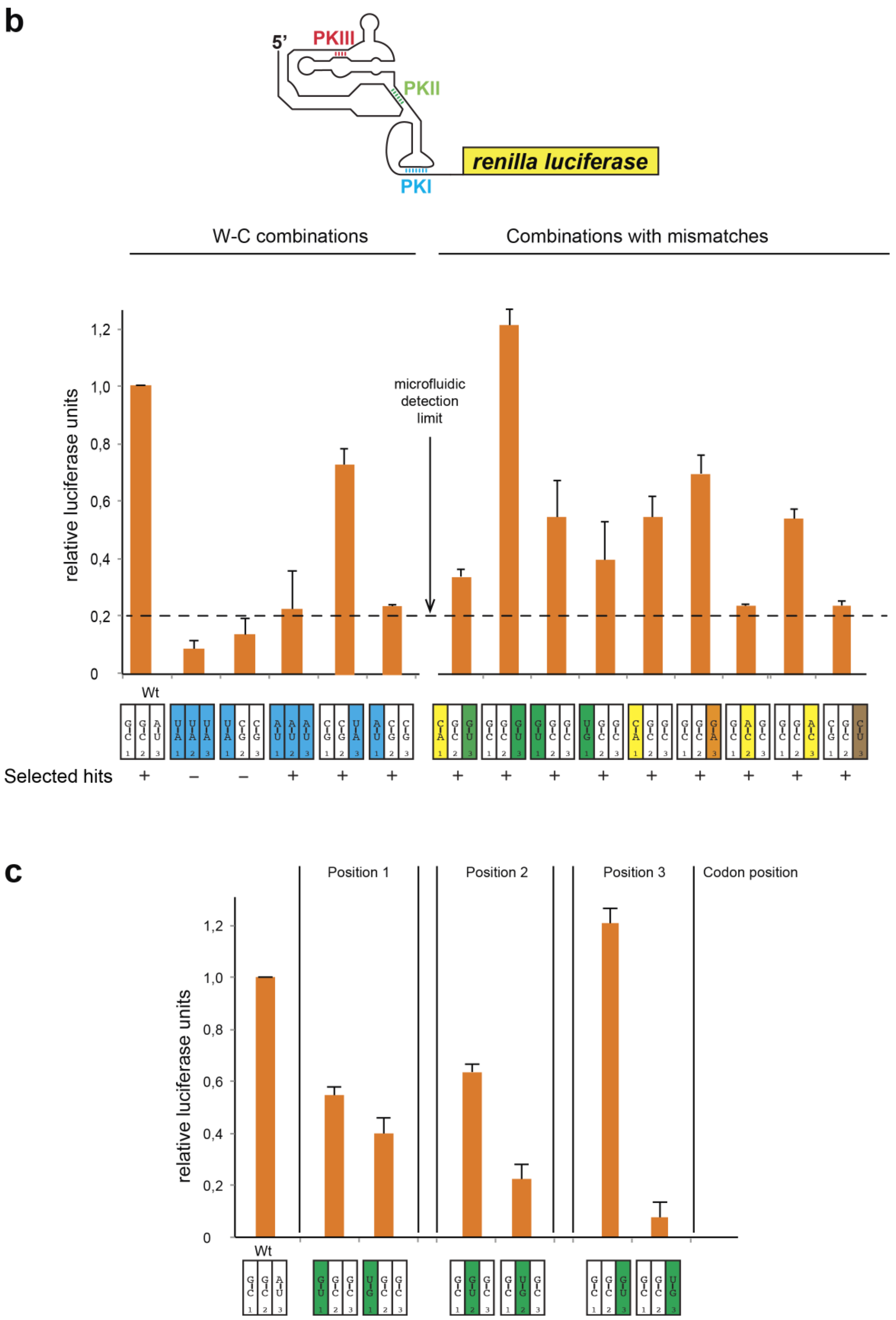

b
5'
PKIII
PKII
PKI
renilla luciferase
W-C combinations
Combinations with mismatches
relative luciferase units
1,2
1,0
0,8
0,6
0,4
0,2
0
microfluidic detection limit
Wt
Selected hits
+ – – + + + + + + + + + + + +
c
Position 1
Position 2
Position 3
Codon position
relative luciferase units
1,2
1,0
0,8
0,6
0,4
0,2
0
Wt

Figure S7

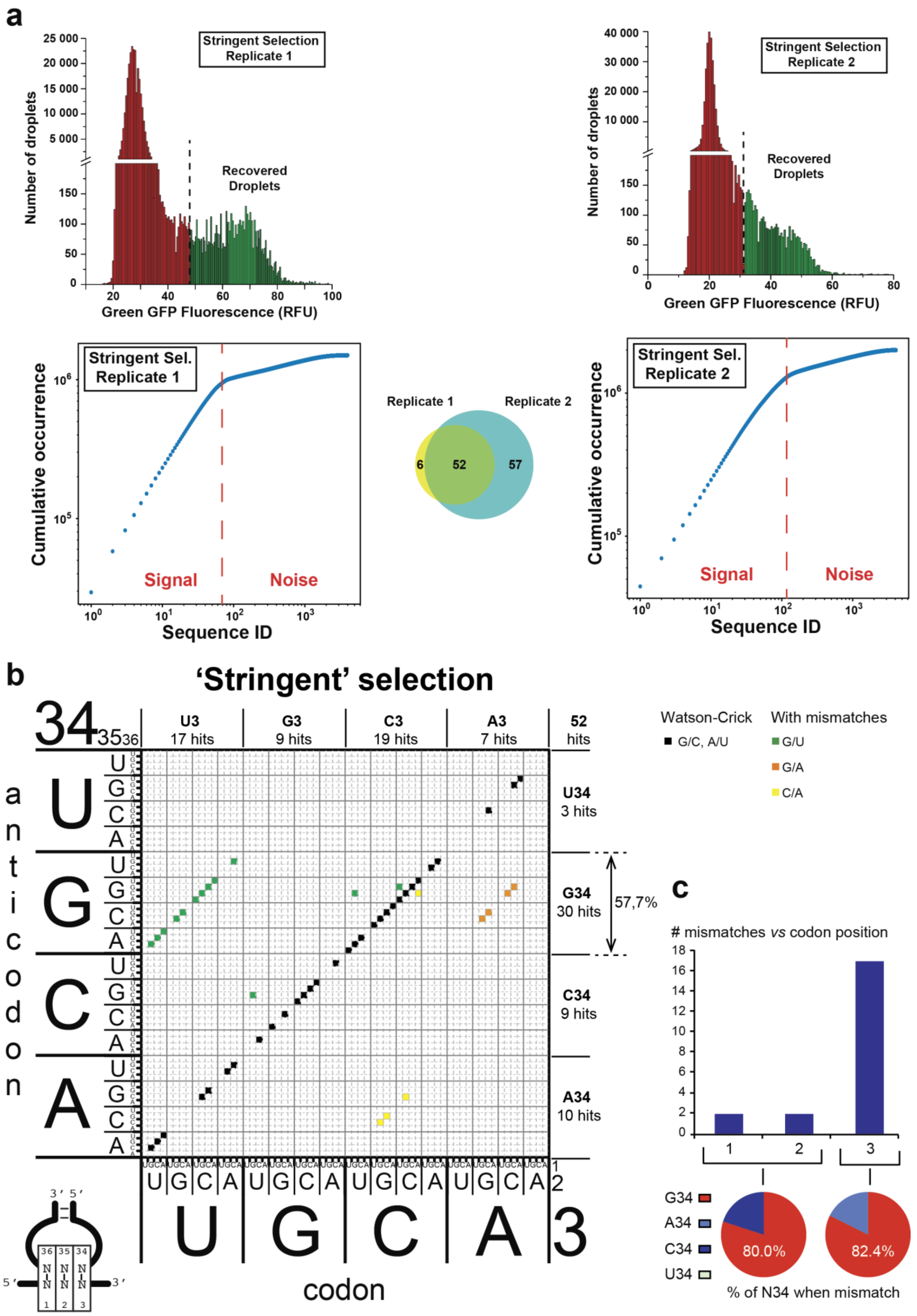

a
Stringent Selection Replicate 1
Number of droplets
25 000
20 000
15 000
10 000
5 000
150
100
50
0
Recovered Droplets
20
40
60
80
100
Green GFP Fluorescence (RFU)
Stringent Selection Replicate 2
40 000
30 000
20 000
10 000
150
100
50
0
Recovered Droplets
0
20
40
60
80
Green GFP Fluorescence (RFU)
Stringent Sel. Replicate 1
Cumulative occurrence
10^6
10^5
Signal
Noise
10^0
10^1
10^2
10^3
Sequence ID
Replicate 1
Replicate 2
6
52
57
Stringent Sel. Replicate 2
Cumulative occurrence
10^6
10^5
Signal
Noise
10^0
10^1
10^2
10^3
Sequence ID
b
'Stringent' selection
34 35 36
U3 17 hits
G3 9 hits
C3 19 hits
A3 7 hits
52 hits
anticodon
U
G
C
A
U34 3 hits
G34 30 hits
57,7%
C34 9 hits
A34 10 hits
UGCA
U G C A U G C A U G C A U G C A
1
2
U G C A
3
codon
3′ 5′
36 35 34
N N N
1 2 3
5′
3′
Watson-Crick
■ G/C, A/U
With mismatches
🟩 G/U
🟧 G/A
🟨 C/A
c
# mismatches vs codon position
18
16
14
12
10
8
6
4
2
0
1
2
3
G34
A34
C34
U34
80.0%
82.4%
% of N34 when mismatch

Figure S8

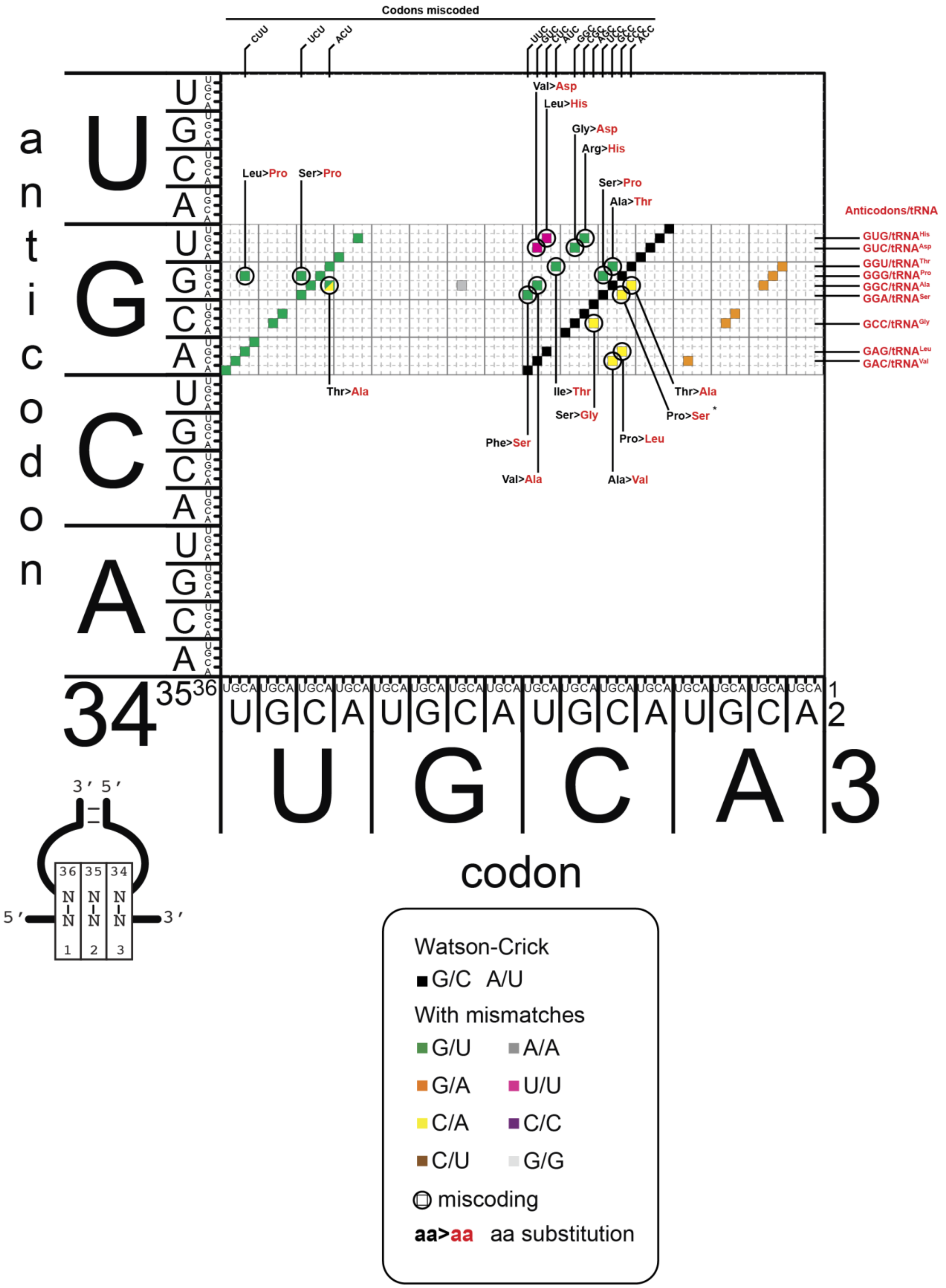

Codons miscoded
CUU
UCU
ACU
UUC
GUC
CUC
AUC
GGC
CGC
AGC
UCC
GCC
CCC
ACC
Val>Asp
Leu>His
Gly>Asp
Arg>His
Leu>Pro
Ser>Pro
Ser>Pro
Ala>Thr
Anticodons/tRNA
GUG/tRNA His
GUC/tRNA Asp
GGU/tRNA Thr
GGG/tRNA Pro
GGC/tRNA Ala
GGA/tRNA Ser
GCC/tRNA Gly
GAG/tRNA Leu
GAC/tRNA Val
Thr>Ala
Ile>Thr
Thr>Ala
Ser>Gly
Pro>Ser *
Phe>Ser
Pro>Leu
Val>Ala
Ala>Val
anticodon
U
G
C
A
34
35
36
1
2
3
codon
3′ 5′
36 35 34
N N N
N N N
1 2 3
5′
3′
Watson-Crick
■ G/C A/U
With mismatches
🟩 G/U
■ A/A
🟧 G/A
■ U/U
🟨 C/A
■ C/C
■ C/U
■ G/G
○ miscoding
aa>aa aa substitution

Figure S9

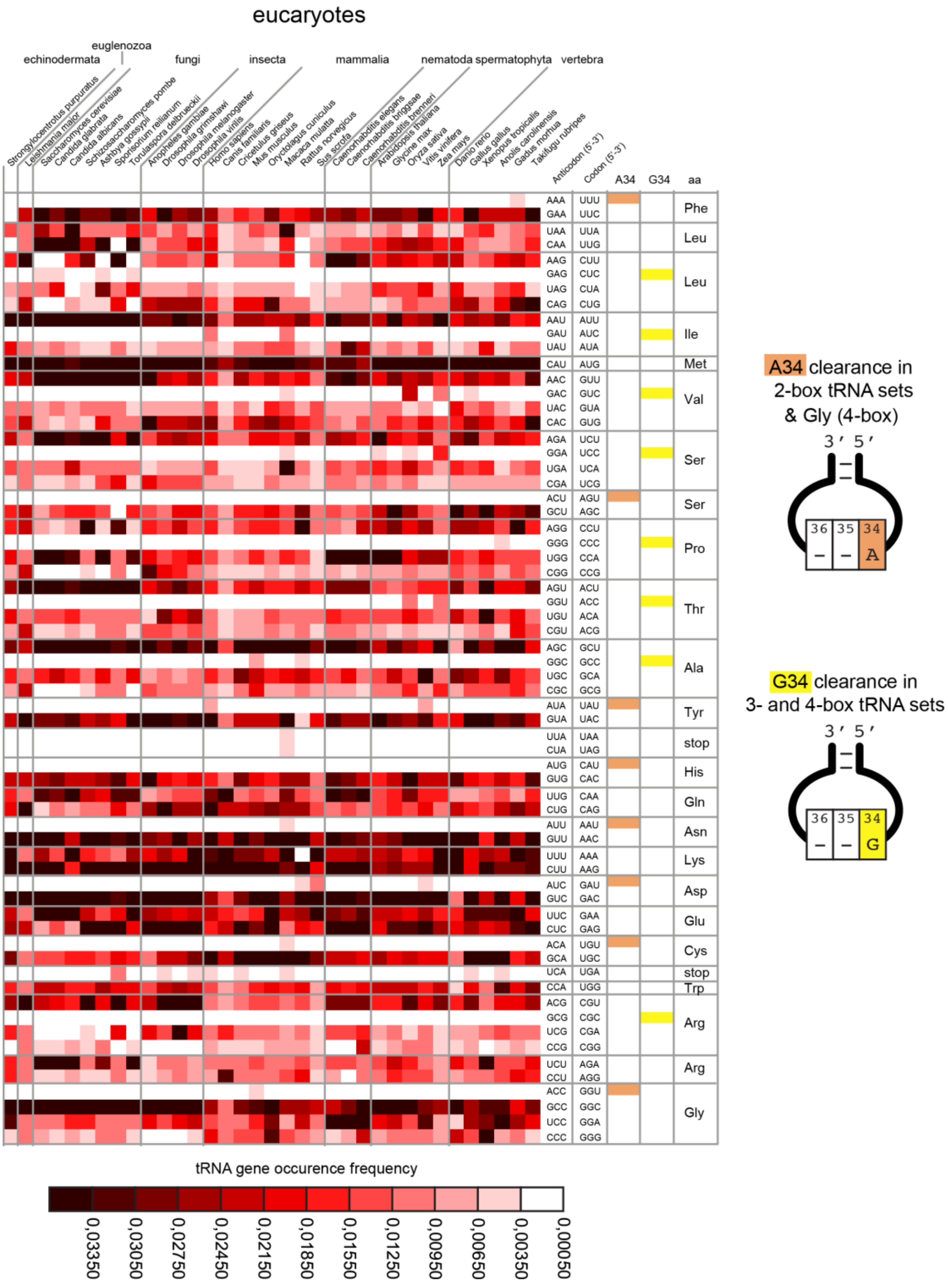
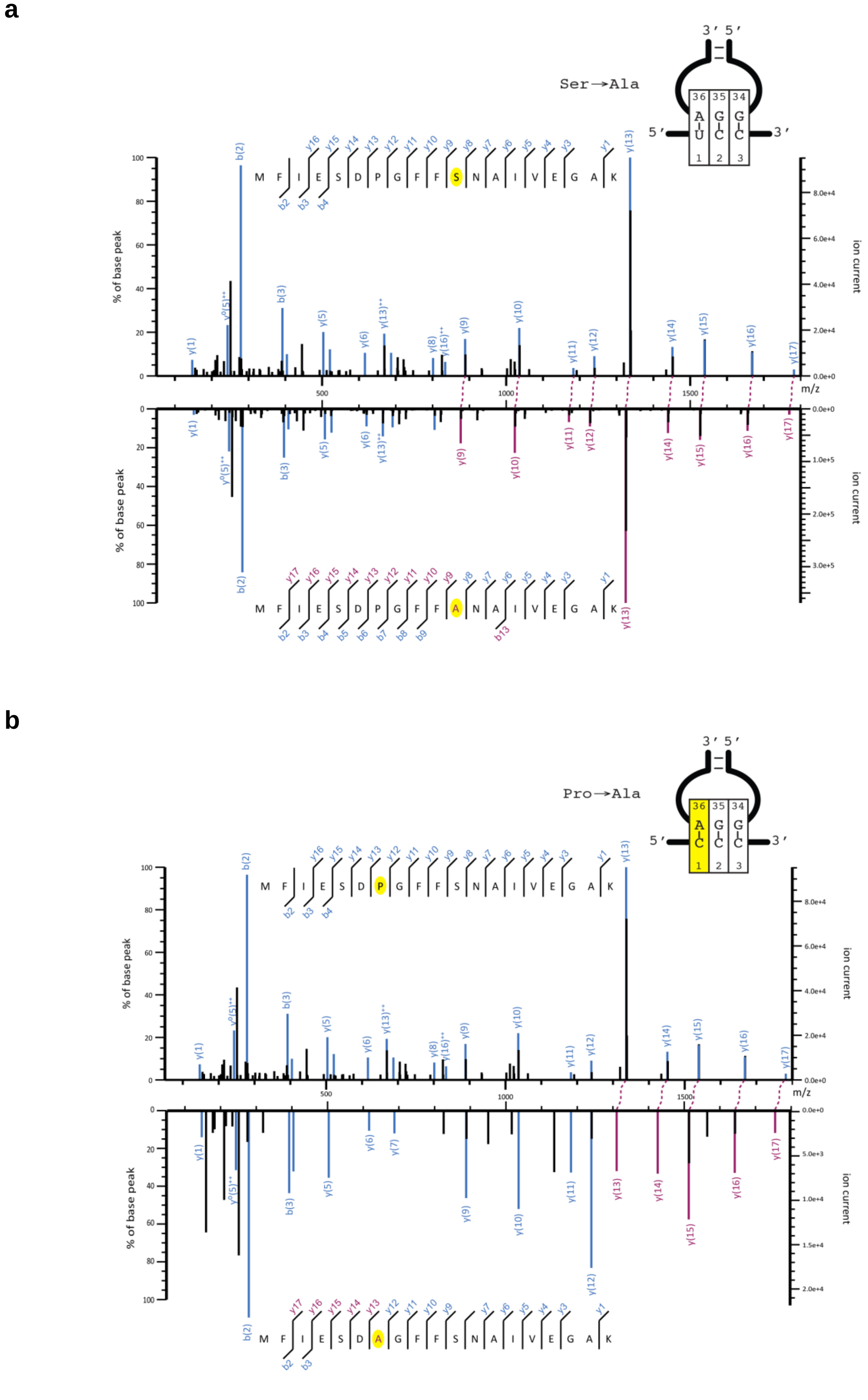

eucaryotes
echinodermata
euglenozoa
fungi
insecta
mammalia
nematoda
spermatophyta
vertebra
Strongylocentrotus purpuratus
Leishmania major
Saccharomyces cerevisiae
Candida glabrata
Candida albicans
Schizosaccharomyces pombe
Ashbya gossypii
Sporisorium reilianum
Torulaspora delbrueckii
Anopheles gambiae
Drosophila grimshawi
Drosophila melanogaster
Drosophila virilis
Homo sapiens
Canis familiaris
Cricetulus griseus
Mus musculus
Oryctolagus cuniculus
Macaca mulatta
Rattus norvegicus
Sus scrofa
Caenorhabditis elegans
Caenorhabditis briggsae
Caenorhabditis brenneri
Arabidopsis thaliana
Glycine max
Oryza sativa
Vitis vinifera
Zea mays
Danio rerio
Gallus gallus
Xenopus tropicalis
Anolis carolinensis
Gadus morhua
Takifugu rubripes
Anticodon (5'-3')
Codon (5'-3')
A34
G34
aa
AAA UUU
GAA UUC
Phe
UAA UUA
CAA UUG
Leu
AAG CUU
GAG CUC
UAG CUA
CAG CUG
Leu
AAU AUU
GAU AUC
UAU AUA
Ile
CAU AUG
Met
AAC GUU
GAC GUC
UAC GUA
CAC GUG
Val
AGA UCU
GGA UCC
UGA UCA
CGA UCG
Ser
ACU AGU
GCU AGC
Ser
AGG CCU
GGG CCC
UGG CCA
CGG CCG
Pro
AGU ACU
GGU ACC
UGU ACA
CGU ACG
Thr
AGC GCU
GGC GCC
UGC GCA
CGC GCG
Ala
AUA UAU
GUA UAC
Tyr
UUA UAA
CUA UAG
stop
AUG CAU
GUG CAC
His
UUG CAA
CUG CAG
Gln
AUU AAU
GUU AAC
Asn
UUU AAA
CUU AAG
Lys
AUC GAU
GUC GAC
Asp
UUC GAA
CUC GAG
Glu
ACA UGU
GCA UGC
Cys
UCA UGA
stop
CCA UGG
Trp
ACG CGU
GCG CGC
UCG CGA
CCG CGG
Arg
UCU AGA
CCU AGG
Arg
ACC GGU
GCC GGC
UCC GGA
CCC GGG
Gly
tRNA gene occurence frequency
0,03350
0,03050
0,02750
0,02450
0,02150
0,01850
0,01550
0,01250
0,00950
0,00650
0,00350
0,00050
A34 clearance in
2-box tRNA sets
& Gly (4-box)
3' 5'
36 35 34
– – A
G34 clearance in
3- and 4-box tRNA sets
3' 5'
36 35 34
– – G

Figure S10


a
Ser→Ala
3′ 5′
36 35 34
A G G
U C C
1 2 3
5′ 3′
M F I E S D P G F F S N A I V E G A K
y16 y15 y14 y13 y12 y11 y10 y9 y8 y7 y6 y5 y4 y3 y1
b2 b3 b4
M F I E S D P G F F A N A I V E G A K
y17 y16 y15 y14 y13 y12 y11 y10 y9 y8 y7 y6 y5 y4 y3 y1
b2 b3 b4 b5 b6 b7 b8 b9 b13
% of base peak
ion current
m/z
0 20 40 60 80 100
500 1000 1500
0.0e+0 2.0e+4 4.0e+4 6.0e+4 8.0e+4
0.0e+0 1.0e+5 2.0e+5 3.0e+5
y(1) y⁰(5)⁺⁺ b(2) b(3) y(5) y(6) y(13)⁺⁺ y(8) y(16)⁺⁺ y(9) y(10) y(11) y(12) y(13) y(14) y(15) y(16) y(17)
b
Pro→Ala
3′ 5′
36 35 34
A G G
C C C
1 2 3
5′ 3′
M F I E S D P G F F S N A I V E G A K
y16 y15 y14 y13 y12 y11 y10 y9 y8 y7 y6 y5 y4 y3 y1
b2 b3 b4
M F I E S D A G F F S N A I V E G A K
y17 y16 y15 y14 y13 y12 y11 y10 y9 y7 y6 y5 y4 y3 y1
b2 b3
% of base peak
ion current
m/z
0 20 40 60 80 100
500 1000 1500
0.0e+0 2.0e+4 4.0e+4 6.0e+4 8.0e+4
0.0e+0 5.0e+3 1.0e+4 1.5e+4 2.0e+4
y(1) y⁰(5)⁺⁺ b(2) b(3) y(5) y(6) y(7) y(13)⁺⁺ y(8) y(16)⁺⁺ y(9) y(10) y(11) y(12) y(13) y(14) y(15) y(16) y(17)