## Supplementary material for:

# New families of human regulatory RNA structures identified by comparative analysis of vertebrate genomes

Brian J. Parker*, Ida Moltke, Adam Roth, Stefan Washietl, Jiayu Wen, Manolis Kellis, Ronald Breaker, and Jakob Skou Pedersen*.
*Corresponding authors

Supplementary figures: .......... 3
Figure S1: 29 mammals and two vertebrate out-group species used for family detection. .......... 3
Figure S2: 41-way alignment. .......... 4
Figure S3: Family mean length distribution. .......... 5
Figure S4: Genomic distribution of initial EvoFold prediction set (excluding protein-coding regions). .......... 6
Figure S5: Genomic distribution of family members. .......... 7
Figure S6: *MALAT1* family. .......... 8
Figure S7: Thermodynamic analysis of structure families using RNAz. .......... 9
Figure S8: Correlation of expression between family members. .......... 10
Figure S9: Substitution evidence for *MAT2A* 3′UTR family, hairpin A. .......... 11
Figure S10: Substitution evidence for *MAT2A* 3′UTR family, hairpin B. .......... 12
Figure S11: Substitution evidence for *MAT2A* 3′UTR family, hairpin C. .......... 13
Figure S12: Substitution evidence for *MAT2A* 3′UTR family, hairpin D. .......... 14
Figure S13: Substitution evidence for *MAT2A* 3′UTR family, hairpin E. .......... 15
Figure S14: Substitution evidence for *MAT2A* 3′UTR family, hairpin F. .......... 16
Figure S15: Tissue-specific ratios of *MAT2A* expression comparing upstream and downstream regions of the 3′UTR. .......... 17
Figure S16: RNA constructs used for in-line probing analyses (5′ to 3′). .......... 18
Figure S17: In-line probing analysis of 186 *MAT2A*. .......... 19
Figure S18: In-line probing analysis of 411 *MAT2A*. .......... 20
Figure S19: In-line probing analysis of 358 *MAT2A*. .......... 21
Figure S20: In-line probing analysis of 268 *MAT2A*. .......... 22
Figure S21: Levels of agreement between computationally and experimentally derived *MAT2A* hairpin secondary structure predictions. .......... 24
Figure S22: RNA-seq transcript evidence for *CLCN5* putative microRNA across cell lines. .......... 26
Figure S23: Selected RNA-seq tracks for family GW45. .......... 27
Figure S24: *XIST* family (EvoFam id: GWP786). .......... 28
Figure S25: Difference between the two substitution counting approaches. .......... 30
Supplementary tables: .......... 31
Table S1: Top 10 detected GW families. .......... 31
Table S2: Data set element counts. .......... 32
Table S3: Overlap counts of datasets. .......... 32
Table S4: Known cis-regulatory elements detected in families. .......... 33
Table S5: Recall rates on families of structural RNAs inferred from database annotations. .......... 34
Table S6: Recall rates on families of structural RNAs extracted from the literature. .......... 35
Table S7: Immunity-related families. .......... 36
Table S8: Ion channel gene-related families. .......... 38
Table S9: Members of the ubiquitin-system and immunity enriched family UTRP37/GWP66. .......... 39

Table S10: Long hairpins clustering with known miRNAs in GW set. 40
Table S11: Distributions of mature read lengths in HeLa and MB-MDA231 breast adenocarcinoma for *CLCN5* putative microRNA. 41
Table S12: GO enrichment of lincRNAs. 42
Supplementary results: 43
Other uncharacterized families of hairpins with strong independent evidence: 43
*XIST family GWP786:* 43
*Ubiquitin system-enriched family UTRP37/GWP66:* 43
*BCL11A/B family UTRP2:* 44
*Ion-channel family GW59:* 44
*Alternative splicing putative family GWP413:* 44
Supplementary methods: 45
Genome-wide EvoFold screen for structural RNAs 45
Similarity measure between models of structural RNAs 45
Graph-based family definition 48
Substitution evidence evaluation (EvoP test) 48
Alignment filtering 49
Thermodynamic analysis of structure families with RNAz 50
Computational validation 50
Expression analyses 52
References: 53

## Supplementary data files.

Data file S1: Full set of 220 GW (filtered) family predictions is available in supplementary file: “Parker_et_al_supp_data_file_1.tab” (data format is documented in initial comment lines).
Data file S2: Set of GW (unfiltered) intergenic members overlapping protein-coding potential predictions by Exoniphy in file “Parker_et_al_supp_data_file_2.tab”.
Data file S3: Structural similarity within GW (filtered) families in file “Parker_et_al_supp_data_file_3.tab”**.**

**Note:** All genomic coordinates are relative to the Human March 2006 (hg18) assembly.

## Supplementary figures:

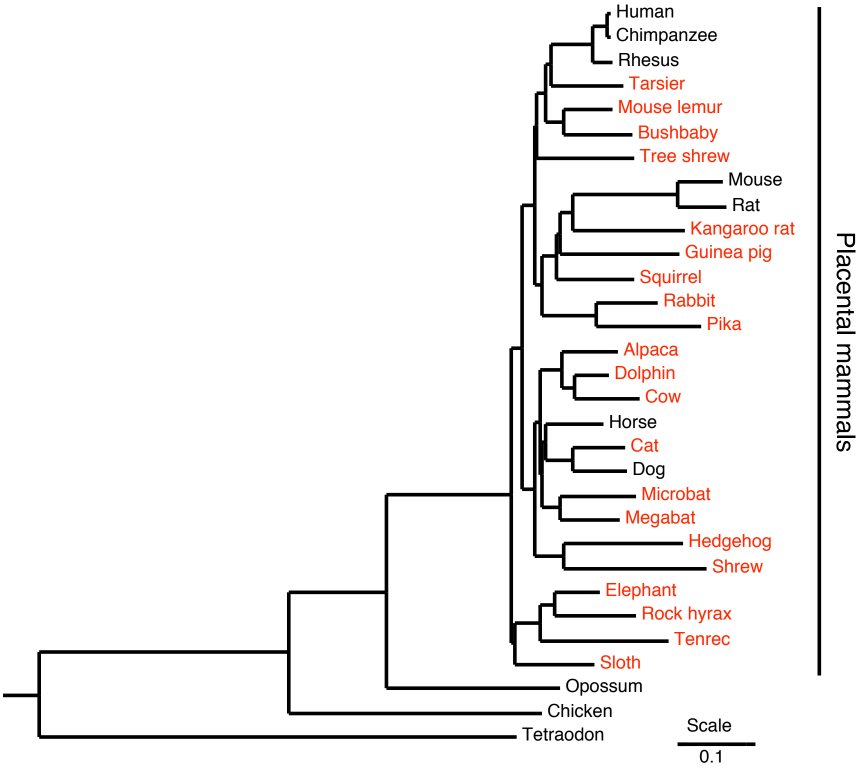


**Figure S1: 29 mammals and two vertebrate out-group species used for family detection.** Species sequenced at 2x low coverage are in red.

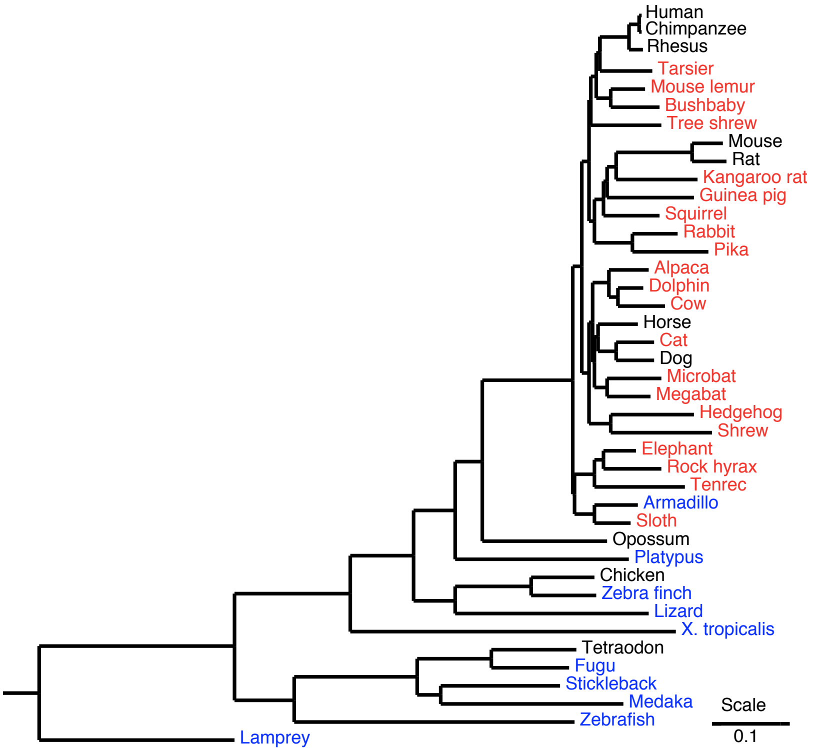


**Figure S2: 41-way alignment.**
Full set of vertebrate species used for both family detection and independent test set. Blue: species not used for inference; Red: species used for inference sequenced at 2x low coverage.

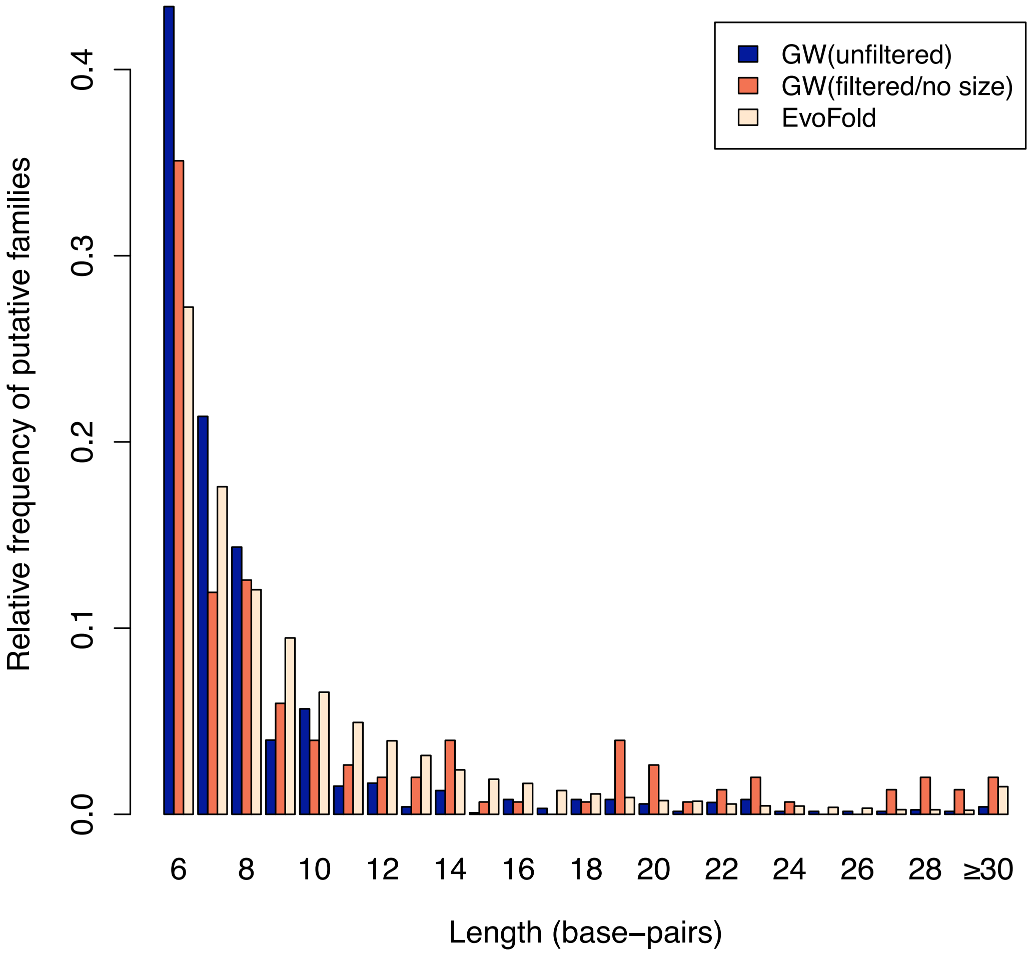


**Figure S3: Family mean length distribution.**
Distribution of family mean length in base-pairs for unfiltered GW set (blue), filtered (high-confidence) GW set (orange), EvoFold background set (white). The GW family distributions show a relative increase in very short (≤8 bp) hairpins. The filtered GW set (excluding filtering on size) shows a skew towards longer structures (18-23 bp) compared with the EvoFold background. Note that structures < 6 bp are removed from the pipeline.

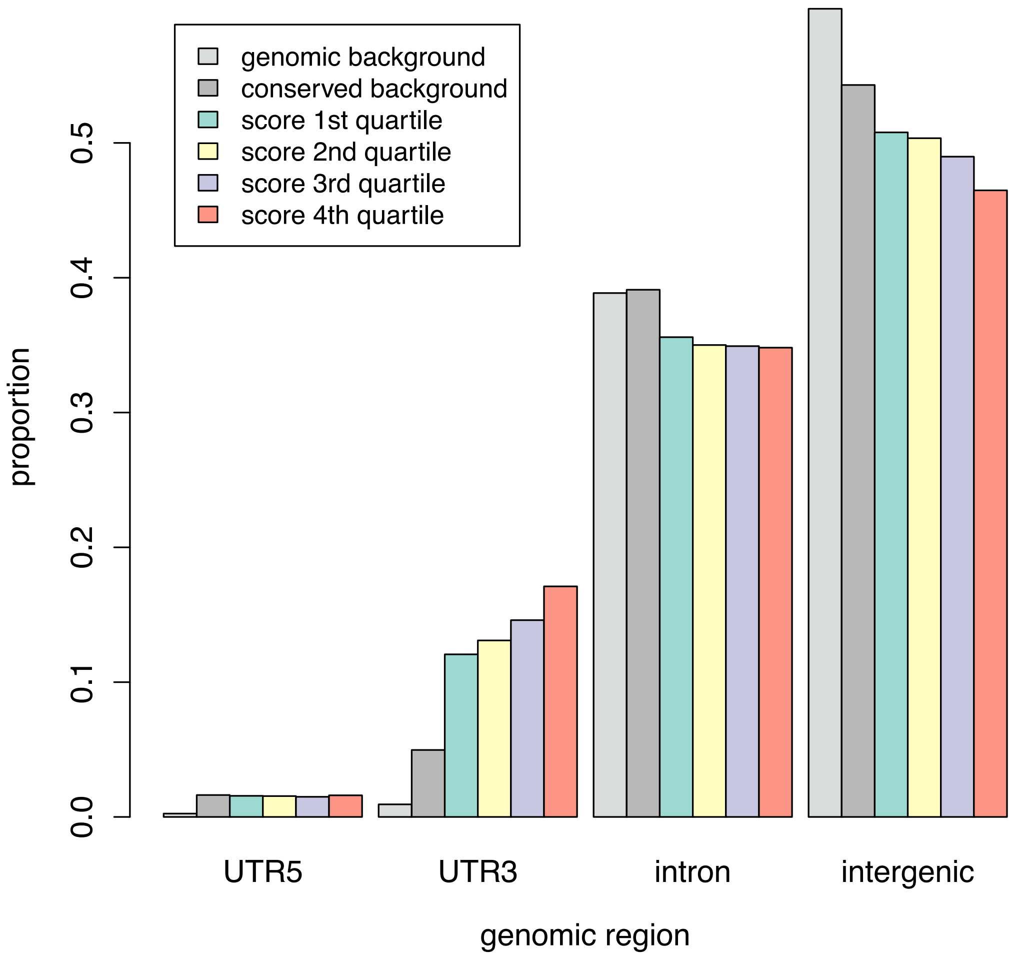


**Figure S4: Genomic distribution of initial EvoFold prediction set (excluding protein-coding regions).**

The distribution is shown for each quartile of predictions based on the EvoFold log odds score. Higher score (higher confidence) predictions are enriched in UTRs, consistent with there being many true cis-regulatory structures in these regions. For comparison, the genomic distribution of the conserved regions used as initial input to the pipeline are shown, along with the overall genomic region proportions.

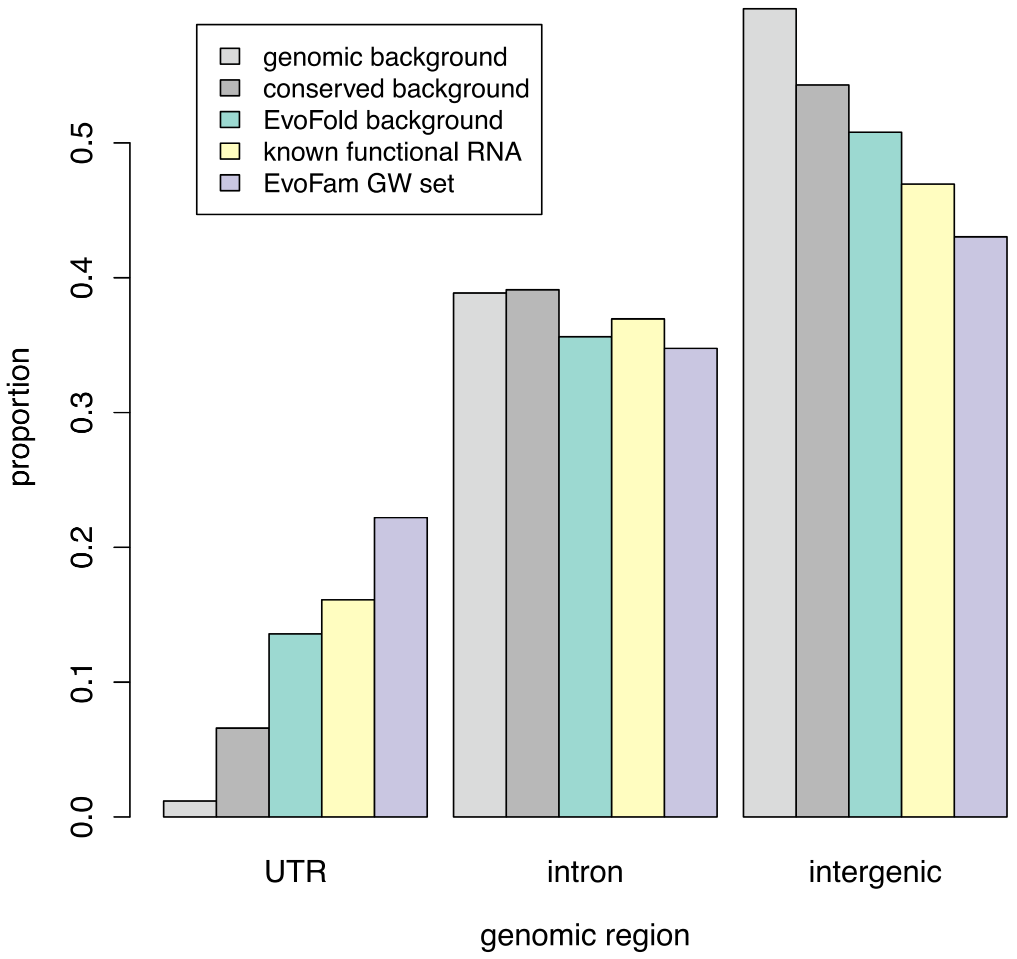
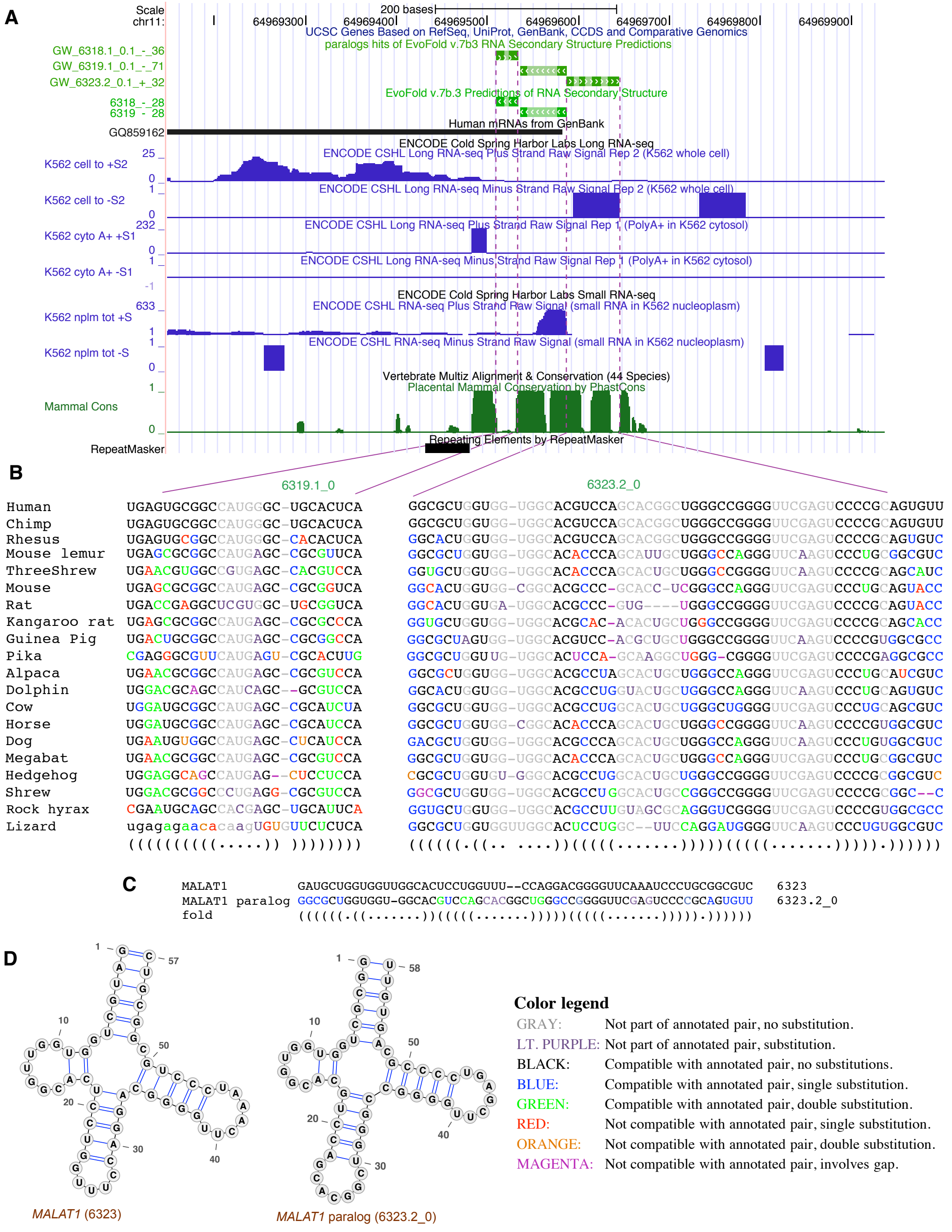


**Figure S5: Genomic distribution of family members.**
The genomic distribution of the high-confidence GW set is compared with both the initial EvoFold set, conserved regions, and with a set of known functional RNAs (see Methods). The overall genomic region distribution is also shown. The family members are enriched for UTR, consistent with existence of many cis-regulatory families.


A
Scale
chr11:
200 bases
64969300
64969400
64969500
64969600
64969700
64969800
64969900
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
paralogs hits of EvoFold v.7b3 RNA Secondary Structure Predictions
GW_6318.1_0.1_-_36
GW_6319.1_0.1_-_71
GW_6323.2_0.1_+_32
EvoFold v.7b.3 Predictions of RNA Secondary Structure
6318_-_28
6319_-_28
Human mRNAs from GenBank
GQ859162
ENCODE Cold Spring Harbor Labs Long RNA-seq
ENCODE CSHL Long RNA-seq Plus Strand Raw Signal Rep 2 (K562 whole cell)
K562 cell to +S2
25
0
ENCODE CSHL Long RNA-seq Minus Strand Raw Signal Rep 2 (K562 whole cell)
K562 cell to -S2
1
0
ENCODE CSHL Long RNA-seq Plus Strand Raw Signal Rep 1 (PolyA+ in K562 cytosol)
K562 cyto A+ +S1
232
0
ENCODE CSHL Long RNA-seq Minus Strand Raw Signal Rep 1 (PolyA+ in K562 cytosol)
K562 cyto A+ -S1
1
-1
ENCODE Cold Spring Harbor Labs Small RNA-seq
ENCODE CSHL RNA-seq Plus Strand Raw Signal (small RNA in K562 nucleoplasm)
K562 nplm tot +S
633
1
ENCODE CSHL RNA-seq Minus Strand Raw Signal (small RNA in K562 nucleoplasm)
K562 nplm tot -S
1
0
Vertebrate Multiz Alignment & Conservation (44 Species)
Placental Mammal Conservation by PhastCons
Mammal Cons
1
0
Repeating Elements by RepeatMasker
RepeatMasker
B
6319.1_0
6323.2_0
C
D
MALAT1 (6323)
MALAT1 paralog (6323.2_0)
1
10
20
30
40
50
57
58

```
Human         UGAGUGCGGCCAUGGGC-UGCACUCA    GGCGCUGGUGG-UGGCACGUCCAGCACGGCUGGGCCGGGGUUCGAGUCCCCGCAGUGUU
Chimp         UGAGUGCGGCCAUGGGC-UGCACUCA    GGCGCUGGUGG-UGGCACGUCCAGCACGGCUGGGCCGGGGUUCGAGUCCCCGCAGUGUU
Rhesus        UGAGUGCGGCCAUGGGC-CACACUCA    GGCACUGGUGG-UGGCACGUCCAGCACGGCUGGGCCGGGGUUCGAGUCCCCGCAGUGUC
Mouse lemur   UGAGCGCGGCCAUGAGC-CGCGUUCA    GGCGCUGGUGG-UGGCACACCCAGCAUUGCUGGGCCAGGGUUCAAGUCCCUGCGGCGUC
ThreeShrew    UGAACGUGGCCGUGAGC-CACGUCCA    GGUGCUGGUGG-UGGCACACCCAGCACUGCUGGGCCGGGGUUCAAGUCCCCGCAGCAUC
Mouse         UGAGCGCGGCCAUGAGC-CGCGGUCA    GGCACUGGUGG-CGGCACGCCC-GCACC-UCGGGCCAGGGUUCGAGUCCCUGCAGUACC
Rat           UGACCGAGGCUCGUGGC-UGCGGUCA    GGCACUGGUGA-UGGCACGCCC-GUG----UGGGCCGGGGUUCGAGUCCCCGCAGUACC
Kangaroo rat  UGAGCGCGGCCAUGAGC-CGCGCCCA    GGUGCUGGUGG-UGGCACGCAC-ACACUGCUGGGCCGGGGUUCGAGUCCCCGCAGCACC
Guinea Pig    UGACUGCGGCCAUGAGC-CGCGGCCA    GGCGCUAGUGG-UGGCACGUCC-ACGCUGCUGGGCCGGGGUUCAAGUCCCCGUGGCGCC
Pika          CGAGGGCGUUCAUGAGU-CGCACUUG    GGCGCUGGUUG-UGGCACUCCA-GCAAGGCUGGG-CGGGGUUCGAGUCCCCGAGGCGCC
Alpaca        UGAACGCGGCCAUGAGC-CGCGUCCA    GGCGCUGGUGG-UGGCACGCCUAGCACUGCUGGGCCAGGGUUCGAGUCCCUGCAUCGUC
Dolphin       UGGACGCAGCCAUCAGC--GCGUCCA    GGCACUGGUGG-UGGCACGCCUGGUACUGCUGGGCCAGGGUUCAAGUCCCUGCAGUGUC
Cow           UGGAUGCGGCCAUGAGC-CGCAUCUA    GGCGCUGGUGG-UGGCACGCCUGGCACUGCUGGGCUGGGGUUCGAGUCCCUGCAGCGUC
Horse         UGGAUGCGGCCAUGAGC-CGCAUCCA    GGCGCUGGUGG-CGGCACACCCAGCACUGCUGGGCCGGGGUUCAAGUCCCCGUGGCGUC
Dog           UGAAUGUGGCCAUGAGC-CUCAUCCA    GACGCUGGUGG-UGGCACGCCCAGCACUGCUGGGCCAGGGUUCAAGUCCCUGUGGCGUC
Megabat       UGAACGCGGCCAUGAGC-CGCGUCCA    GGCGCUGGUGG-UGGCACACCCAGCACUGCUGGGCCAGGGUUCGAGUCCCUGCGGCGUC
Hedgehog      UGGAGGCAGCCAUGAG--CUCCUCCA    CGCGCUGGUGU-GGGCACGCCUGGCACUGCUGGGCCGGGGUUCGAGUCCCCGCGGCGUC
Shrew         UGGACGCGGCCCUGAGG-CGCGUCCA    GGCGCUGGUGG-UGGCACGCCUGGCACUGCCGGGCCGGGGUUCGAGUCCCCGCGGC--C
Rock hyrax    CGAAUGCAGCCACGAGC-UGCAUUCA    GGUGCUGGUGG-UGGCACGCCUUGUAGCGCAGGGUCGGGGUUCGAGUCCCCGUGGCGCC
Lizard        ugagagaacacaagUGUGUUCUCUCA    GGCGCUGGUGGUUGGCACUCCUGGC--UUCCAGGAUGGGGUUCAAGUCCCUGUGGCGUC
              ((((((((((.....)) ))))))))    ((((((.((.. ....))(((((.......)))))(((((.......))))).))))))
```

```
MALAT1          GAUGCUGGUGGUUGGCACUCCUGGUUU--CCAGGACGGGGUUCAAAUCCCUGCGGCGUC   6323
MALAT1 paralog  GGCGCUGGUGGU-GGCACGUCCAGCACGGCUGGCCGGGGUUCGAGUCCCCGCAGUGUU   6323.2_0
fold            ((((((.((.......))(((((.......)))))(((((.......))))).))))))
```

**Color legend**

| | |
|---|---|
| GRAY: | Not part of annotated pair, no substitution. |
| LT. PURPLE: | Not part of annotated pair, substitution. |
| BLACK: | Compatible with annotated pair, no substitutions. |
| BLUE: | Compatible with annotated pair, single substitution. |
| GREEN: | Compatible with annotated pair, double substitution. |
| RED: | Not compatible with annotated pair, single substitution. |
| ORANGE: | Not compatible with annotated pair, double substitution. |
| MAGENTA: | Not compatible with annotated pair, involves gap. |

**Figure S6: *MALAT1* family.**
(A) Detected paralogous member to *MALAT1* is shown, located downstream of ncRNA gene *NEAT1 (*transcript *MEN β)*. EST and RNA-seq data (Birney et al. 2007) demonstrate the end cleavage by RNaseP (Wilusz and Spector 2010). (B) Subset of species alignment showing substitutions. (C) Alignment of the human sequences of the two cloverleaf members of the family. (D) Structure diagrams.

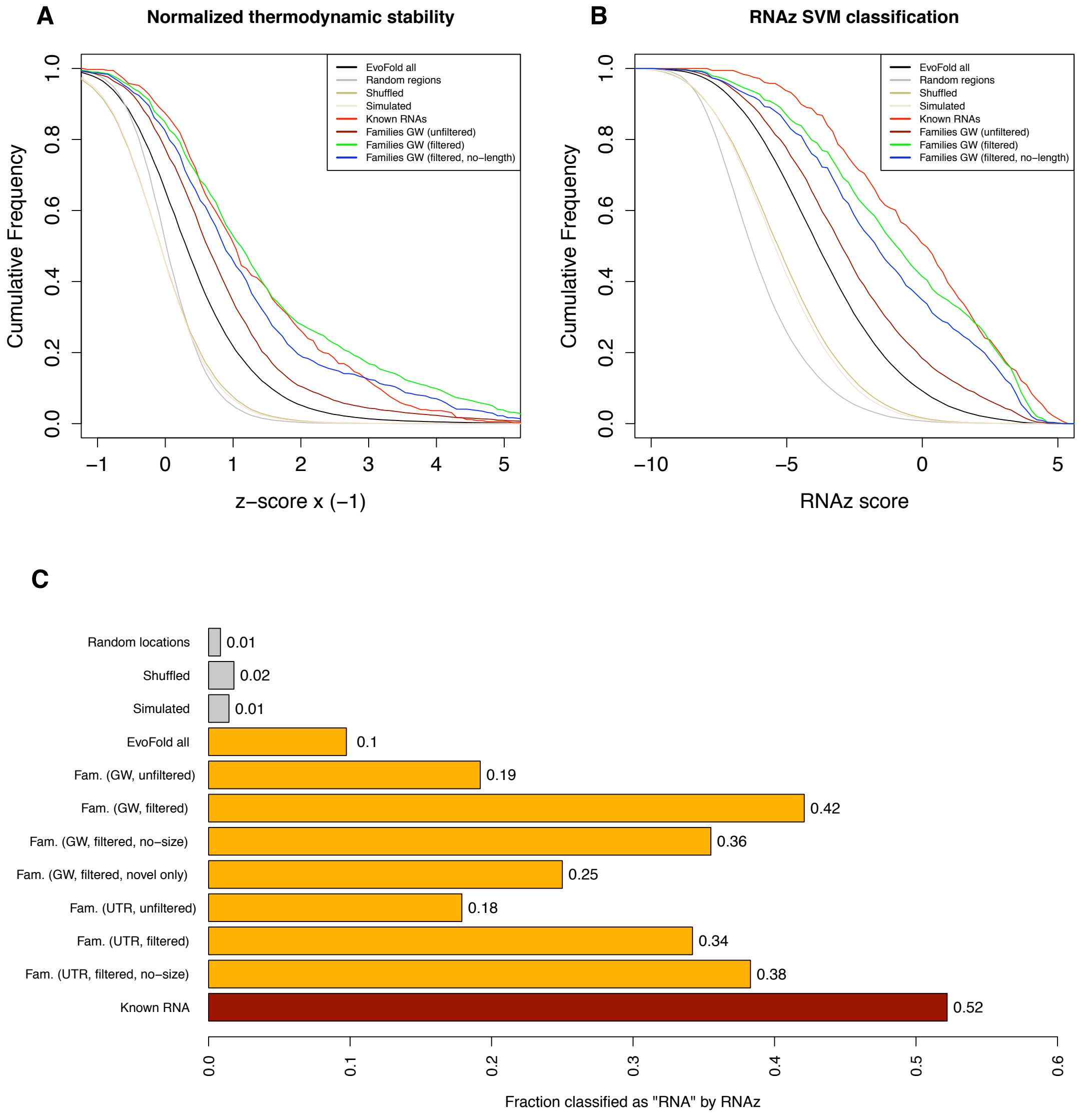


**Figure S7: Thermodynamic analysis of structure families using RNAz.**
(A) Cumulative frequency distribution of normalized thermodynamic stability z-scores. (B) Cumulative frequency distribution of RNAz classification scores. (C) Fraction of structures that were classified as “structural RNA” by the RNAz. For a more detailed description of the sets and metrics refer to Supplementary Text.

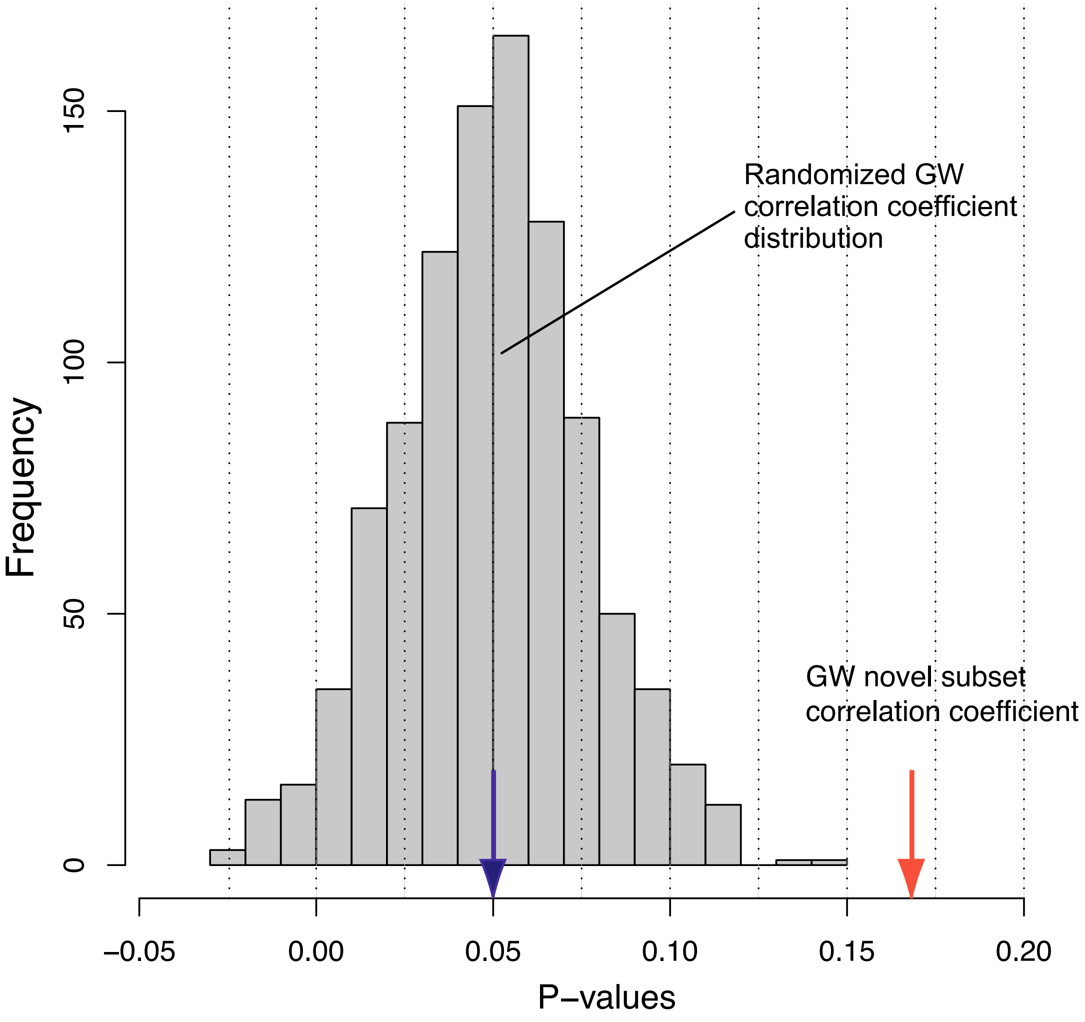


**Figure S8: Correlation of expression between family members.**
Mean pairwise correlation of expression within families, across 16 tissues (Illumina Inc. 2010), compared with randomized set. Mean Pearson correlation coefficient was computed between all pairs of novel members of GW (filtered) families (excluding members with no substantial expression across all tissues; duplicate members within one gene within a family removed), and compared with the distribution of 1000 shuffled family definitions. The unshuffled set shows significantly higher mean correlation suggesting a degree of co-regulation between family members. This result was replicated on an independent multi-tissue RNA-seq dataset (Wang et al. 2008) with correlation coefficients of 0.158 and 0.073 for GW set and random background respectively (p-value=3e-4).

```
position               10        20        30        40        50        60
Human         GCCUUUUUUCCCCAGACUUGUUGGCGUAGGCUACAGAGAAGCCUUCAAGCUCUGAGGGAAAGGGC
Chimp         GCCUUUUUUCCCCAGACUUGUUGGCGUAGGCUACAGAGAAGCCUUCAAGCUCUGAGGGAAAGGGC
Rhesus        GCCUUUUUUCCCCAGACUUGUUGGCGUAGGCUACAGAGAAGCCUUCAAGCUCUGAGGGAAAGGGC
Tarsier       GACUUUUCUCUCCAGACUUGUUGGCGUAGGCUACAGAGAAGCCUUCAAGCUCUGAGGGAAAGGGA
Mouse lemur   GCCUUUUUUCCCCAGACUUGUUGGCGUAGGCUACAGAGAAGCCUUCAAGCUCUGAGGGAAAGGGC
TreeShrew     GCCUUUUUUCCCCAGACUUGUUGGCGUAGGCUACAGAGAAGCCUUCAAGCUCUGAGGGAAAGGGC
Mouse         GCCUUUUUUCCCCAGACUUGUUGGCGUAGGUUACAGAGAAGCCUUCAAGCUCUGAGGGAAAGGGC
Rat           GCCUUUAUUCCCCAGACAUGUUGGCUUAGGUUACAGAGAAGCCUUCAAGCUCUGAGGGAAAGGGC
Kangaroo rat  GCCUUUUUUCCCCAGACUUGUUGGCGUAGGUUACAGAGAAGCCUUCAAGCUCUGAGGGAAAGGGC
Guinea Pig    GCCUUUUUUCCCCAGACUUGUUGGCGUAGGCUACAGAGAAGCCUUCAAGCUCUGAGGGAAAGGGC
Rabbit        GCCUUUUUUCCCCAGACUUGUUGGCGUAGGCUACAGAGAAGCCUUCAAGCUCUGAGGGAAAGGGC
Pika          GCCUUUUUUCCCCAGACUUGUUGGCGUAGGCUACAGAGAAGCCUUCAAGCUCUGAGGGAAAGGGC
Alpaca        GCCUUUUUUCCCCAGACUUGUUGGCGUAGGCUACAGAGAAGCCUUCAAGCUCUGAGGGAAAGGGC
Dolphin       GCCUUUUUUCCCCAGACUUGUUGGCGUAGGCUACAGAGAAGCCUUCAAGCUCUGAGGGAAAGGGC
Cow           GCCUUUUUUCCCCAGACUUGUUGGCGUAGGCUACAGAGAAGCCUUCAAGCUCUGCGGGAAAGGGC
Horse         GCCUUUUUUCCCCAGACUUGUUGGCGUAGGCUACAGAGAAGCCUUCAAGCUCUGAGGGAAAGGGC
Cat           GCCUUUUUUCCCCAGACUUGUUGGCGUAGGCUACAGAGAAGCCUUCAAGCUCUGAGGGAAAGGGC
Dog           GCCUUUUUUCCCCAGACUUGUUGGCGUAGGCUACAGAGAAGCCUUCAAGCUCUGAGGGAAAGGGC
Microbat      GCCUUUUUUCCCCAGACUUGUUGGCGUAGGCUACAGAGAAGCCUUCAAGUUCUGAGGGAAAGGGC
Megabat       GCCUUUUUCCCCCAGACUUGUUGGCGUAGGCUACAGAGAAGCCUUCAAGCUCUGAGGGAAAGGGC
Hedgehog      GCCUUUUUUCCCCAGACUUGUUGGCGUAGGCUACAGAGAAGCCUUCAAGCUCUGAGGGAAAGGGC
Elephant      GCCUUUUUUCCCCAGACUUGUUGGCGUAGGCUACAGAGAAGCCUUCAAGCUCUGAGGGAAAGGGC
Rock hyrax    GCCUUUUUUCCCCAGACUUGUUGGCGUAGGCUACAGAGAAGCCUUCAAGCUCUGAGGGAAAGGGC
Tenrec        GCCUUUUCUCCCCAGACUUGCUGGCGUGGGCUACAGAGAAGCCUGCAAGCUCUGAGGGAAAGGCC
Armadillo     GCCUUUUUUCCCCAGACUUGUUGGCGUAGGCUACAGAGAAGCUUUCAAGCUCUGAGGGAAACGGC
Sloth         GCCUUUUUUCCCCAGACUUGUUGGCGUAGGCUACAGAGAAGCCUUCAAGCUCUGAGGGAAAGGGC
Opossum       GGCUUUUUUCCCCAGACUUGUUGGCGUAGGUUACAGAGAAGCCUUCAAGCUCUGAGGGAAAGGGC
Platypus      GGCUUCUUUCCCCAGACUUGUUGGCGUAGGUUACAGAGAAGCCUUCAAGCUCUGCGGGAAAGGGC
Lizard        GGCUUCUUUCCCCAGACCUGUUGGUGUACUUUACAGAGAAACCUCCAGGCUCUGUGGGAAAGAAC
Fugu          gcuuuuuccccccaggcCUGUUGGUGUAUACUACAGAGAAGCCUCCAGGGUUCCGGGGGAAAUGC
Tetraodon     GC-UUUUCCCCUCAGGCCUGUUGGUGUAUACUACAGAGAAGCCUCCAAGGUUCCGGGGGAAAUGC
Stickleback   GCUUUUUCCCCCCAGGCCUGUUGGUGUAUACUACAGAGAAGCCUCCAAGGUUCCGGGGGAAAUGC
Medaka        GC-UUUUCCCCCUAGGCUUGUUGGUGUAUACUACAGAGAAGCCUCCAAGGUUCCAGGGGAAAUGC
Zebrafish     CCCCUCCUUUCCUUGGGCAGUAGGUGUACAUUACAGAGAAACCUACUCAACCCAAGGGGAGAAGU
fold          ((((...((((.((((..((..(((((.....))......)))..))...)))).))))..))))
pair symbols  abcd   efgh iklm  no  pqrst     ts      rqp  on   mlki hgfe  dcba
                     ↑                                              ↑
```

No change

■ Conserved paired nucleotide

■ Conserved unpaired nucleotide

Changes characteristic of RNA evolution

■ Silent G•U substitution

■ Silent substitution in unpaired base

■ Silent base-preserving double substitution

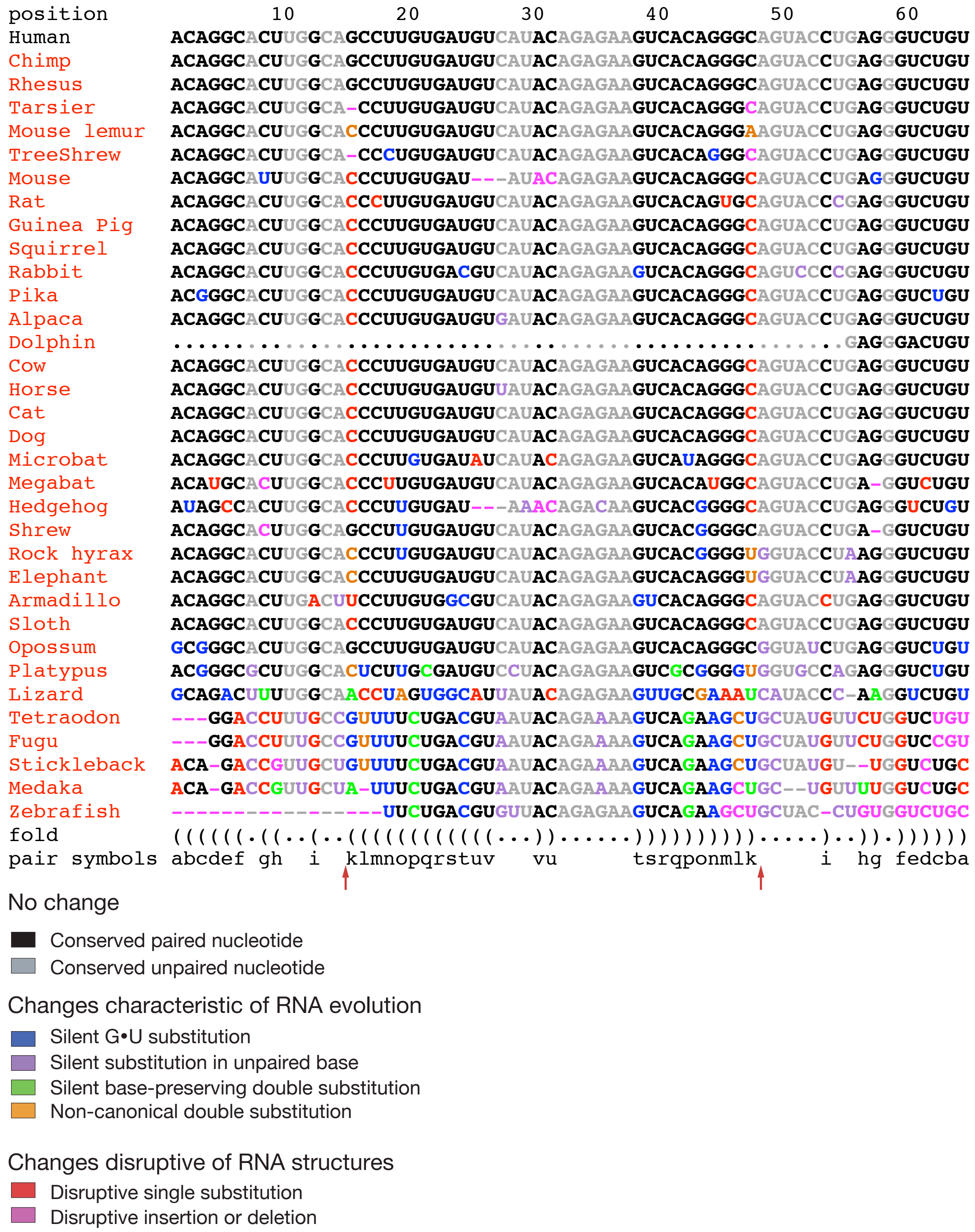
■ Non-canonical double substitution

Changes disruptive of RNA structures

■ Disruptive single substitution

■ Disruptive insertion or deletion

**Figure S9: Substitution evidence for *MAT2A* 3′UTR family, hairpin A.**
Segment of 41-way genomic alignment color-coded according substitution evidence for structure prediction. Species not used for inference are shown in red. Only human was used for inference in this case, since the structure was found by a homology search. The extension of the originally predicted structure, by single sequence energy minimization folding, is indicated with red arrows at the bottom of the alignment.



```
position              10        20        30        40        50        60
Human         ACAGGCACUUGGCAGCCUUGUGAUGUCAUACAGAGAAGUCACAGGGCAGUACCUGAGGGUCUGU
Chimp         ACAGGCACUUGGCAGCCUUGUGAUGUCAUACAGAGAAGUCACAGGGCAGUACCUGAGGGUCUGU
Rhesus        ACAGGCACUUGGCAGCCUUGUGAUGUCAUACAGAGAAGUCACAGGGCAGUACCUGAGGGUCUGU
Tarsier       ACAGGCACUUGGCA-CCUUGUGAUGUCAUACAGAGAAGUCACAGGGCAGUACCUGAGGGUCUGU
Mouse lemur   ACAGGCACUUGGCACCCUUGUGAUGUCAUACAGAGAAGUCACAGGGAAGUACCUGAGGGUCUGU
TreeShrew     ACAGGCACUUGGCA-CCCUGUGAUGUCAUACAGAGAAGUCACAGGGCAGUACCUGAGGGUCUGU
Mouse         ACAGGCAUUUGGCACCCUUGUGAU---AUACAGAGAAGUCACAGGGCAGUACCUGAGGGUCUGU
Rat           ACAGGCACUUGGCACCCUUGUGAUGUCAUACAGAGAAGUCACAGUGCAGUACCCGAGGGUCUGU
Guinea Pig    ACAGGCACUUGGCACCCUUGUGAUGUCAUACAGAGAAGUCACAGGGCAGUACCUGAGGGUCUGU
Squirrel      ACAGGCACUUGGCACCCUUGUGAUGUCAUACAGAGAAGUCACAGGGCAGUACCUGAGGGUCUGU
Rabbit        ACAGGCACUUGGCACCCUUGUGACGUCAUACAGAGAAGUCACAGGGCAGUCCCCGAGGGUCUGU
Pika          ACGGGCACUUGGCACCCUUGUGAUGUCAUACAGAGAAGUCACAGGGCAGUACCUGAGGGUCUGU
Alpaca        ACAGGCACUUGGCACCCUUGUGAUGUGAUACAGAGAAGUCACAGGGCAGUACCUGAGGGUCUGU
Dolphin       ......................................................GAGGGACUGU
Cow           ACAGGCACUUGGCACCCUUGUGAUGUCAUACAGAGAAGUCACAGGGCAGUACCUGAGGGUCUGU
Horse         ACAGGCACUUGGCACCCUUGUGAUGUUAUACAGAGAAGUCACAGGGCAGUACCUGAGGGUCUGU
Cat           ACAGGCACUUGGCACCCUUGUGAUGUCAUACAGAGAAGUCACAGGGCAGUACCUGAGGGUCUGU
Dog           ACAGGCACUUGGCACCCUUGUGAUGUCAUACAGAGAAGUCACAGGGCAGUACCUGAGGGUCUGU
Microbat      ACAGGCACUUGGCACCCUUGUGAUAUCAUACAGAGAAGUCAUAGGGCAGUACCUGAGGGUCUGU
Megabat       ACAUGCACUUGGCACCCUUGUGAUGUCAUACAGAGAAGUCACAUGGCAGUACCUGA-GGUCUGU
Hedgehog      AUAGCCACUUGGCACCCUUGUGAU---AAACAGACAAGUCACGGGGCAGUACCUGAGGGUCUGU
Shrew         ACAGGCACUUGGCAGCCUUGUGAUGUCAUACAGAGAAGUCACGGGGCAGUACCUGA-GGUCUGU
Rock hyrax    ACAGGCACUUGGCACCCUUGUGAUGUCAUACAGAGAAGUCACGGGUGGUACCUAAGGGUCUGU
Elephant      ACAGGCACUUGGCACCCUUGUGAUGUCAUACAGAGAAGUCACAGGGUGGUACCUAAGGGUCUGU
Armadillo     ACAGGCACUUGACUUCCUUGUGGCGUCAUACAGAGAAGUCACAGGGCAGUACCUGAGGGUCUGU
Sloth         ACAGGCACUUGGCACCCUUGUGAUGUCAUACAGAGAAGUCACAGGGCAGUACCUGAGGGUCUGU
Opossum       GCGGGCACUUGGCAGCCUUGUGAUGUCAUACAGAGAAGUCACAGGGCGGUAUCUGAGGGUCUGU
Platypus      ACGGGCGCUUGGCACUCUUGCGAUGUCCUACAGAGAAGUCGCGGGGUGGUGCCAGAGGGUCUGU
Lizard        GCAGACUUUUGGCAACCUAGUGGCAUUAUACAGAGAAGUUGCGAAAUCAUACCC-AAGGUCUGU
Tetraodon     ---GGACCUUUGCCGUUUUCUGACGUAAUACAGAAAAGUCAGAAGCUGCUAUGUUCUGGUCUGU
Fugu          ---GGACCUUUGCCGUUUUCUGACGUAAUACAGAAAAGUCAGAAGCUGCUAUGUUCUGGUCCGU
Stickleback   ACA-GACCGUUGCUGUUUUCUGACGUAAUACAGAAAAGUCAGAAGCUGCUAUGU--UGGUCUGC
Medaka        ACA-GACCGUUGCUA-UUUCUGACGUAAUACAGAAAAGUCAGAAGCUGC--UGUUUUGGUCUGC
Zebrafish     -----------------UUCUGACGUGUUACAGAGAAGUCAGAAGCUGCUAC-CUGUGGUCUGC
fold          ((((((.((..(..((((((((((((...))......)))))))))).....)..)).))))))
pair symbols  abcdef gh  i  klmnopqrstuv   vu      tsrqponmlk     i  hg fedcba
```

No change

■ Conserved paired nucleotide

■ Conserved unpaired nucleotide

Changes characteristic of RNA evolution

■ Silent G•U substitution

■ Silent substitution in unpaired base

■ Silent base-preserving double substitution

■ Non-canonical double substitution

Changes disruptive of RNA structures

■ Disruptive single substitution

■ Disruptive insertion or deletion

**Figure S10: Substitution evidence for *MAT2A* 3′UTR family, hairpin B.** Segment of 41-way genomic alignment color-coded according substitution evidence for structure prediction. Species not used for inference are shown in red. Only human was used for inference in this case, since the structure was found by a homology search. The extension of the originally predicted structure, by single sequence energy minimization folding, is indicated with red arrows at the bottom of the alignment.

```
position               10        20        30        40        50
Human          GCUUUCUGAACAGCUGGUGUAGCUACAGAGAAACCAGCUUCCUUCAGAGAGC
Chimp          GCUUUCUGAACAGCUGGUGUAGCUACAGAGAAACCAGCUUCCUUCAGAGAGC
Rhesus         GCUUUCUGAACAGCUGGUGUAGCUACAGAGAAACCAGCUUCCUUCAGAGAGC
Tarsier        GCUUUCUGAACAGCUGGUGUAGCUACAGAGAAACCAGCUUCCUUCAAAGAGC
Mouse lemur    GCUUUCUGAACAGCUGGUGUAGCUACAGAGAAACCAGCUUCCUUCAGAGAGC
TreeShrew      GCUUUCUGAACAGCUGGUGUAGCUACAGAGAAACCAGCUUCCUUCAGAGAGC
Mouse          GCUUUCUGAACAGCUGGUGUAGCUACAGAGAAACCAGCUUCCUUCAGAGAGC
Rat            GCUUUCUGAACAGCUGGUGUAGCUACAGAGAAACCAGCUUCCUUCAGAGAGC
Guinea Pig     GCUUUCUGAACAGCUGGUGUAGCUACAGAGAAACCAGCUUCCUUCAGAGAGC
Squirrel       GCUUUCUGAACAGCUGGUGUAGCUACAGAGAAACCAGCUUCCUUCAGAGAGC
Rabbit         GCUUUCUGAACAGCUGGUGUGGCUACAGAGAAACCAGCUUCCUUCAGAGAGC
Pika           GCUUUCUGAACGGCUGGUGUAGCUACAGAGAAACCAGCUUUCUUCAGAGAGC
Alpaca         GCUUUCUGAACAGCUGGUGUAGCUACAGAGAAACCAGCUUCCUUCAGAGAGC
Dolphin        UCUUUCUGAGCAGCCAGUAGAUCAACUGAGAAACAAAUUUCCUUCAGAAAGA
Cow            GCUUUCUGGACAGCUGGUGUAGCUACAGAGAAACCAGCUUCCUUCAGAGAGC
Horse          GCUUUCUGAACAGCUGGUGUAGCUACAGAGAAACCAGCUUCCUUCAGAGAGC
Cat            GCUUUCUGAACAGCUGGUGUAGCUACAGAGAAACCAGCUUCCUUCAGAGAGC
Dog            GCUUUCUGAACAGCUGGUGUAGCUACAGAGAAACCAGCUUCCUUCAGAGAGC
Microbat       GCUUUCUGAACAGCUGGUGUAGCUACAGAGAAACCAGCUUCCUUCAGAGAGC
Megabat        GC-UUCUGAACAGCUGGUGUAGCUACAGAGAAACCAGCUUCGUUCAGACAGC
Hedgehog       GCUUUCUGAACAGCUGGUGUAGCUACAGAGAAACCAGCUUCCUUCAGAGAGC
Shrew          GCUUUCUGAGUAGUUGGUGUAGCUACAGAGAAACCAGCUUCCUUCAGAGAGC
Rock hyrax     GCUUUCUGAACAGCUGGUGUAGCUACAGAGAAACCAGCUUCCUUCAGAGAGC
Elephant       GCUUUCUGAACAGCUGGUGUAGCUACAGAGAAACCAGCUUCCUUCAGAGAGC
Armadillo      GCUUUCUGAACAGCUGGUGUAGCUACAGAGAAACCAGCUUCCUUCAGAGAGC
Sloth          GCUUUCUGAACAGCUGGUGUAGCUACAGAGAAACCAGCUUCCUUCAGAGAGC
Opossum        GCUUUCUGAGCAGCUGGUGUAGAUACAGAGAAACCAGCUUUCUUCAGAGAGC
Platypus       GCUUUCUGAACAGCUGGUGUAGCUACAGAGAAACCAGCUCCCUUCAGAGAGC
Lizard         GCGUUCUGAAGUCCUGGUGUAGCUACAGAGAAACCAGUUCUCUUCAGAAUGC
Frog           ACUCUGCAGCUGGCUGGUGUGGCUACAGAGAAACCAGUACAGAGCAGAGGAU
Tetraodon      GCCUUUUGGGUGGUGGGUGUAGCUACAGAGAAACCU--CCCCACC-----GC
Fugu           GCCUUuuggguggugggUGUAGCUACAGAGAAA------CCCACC-----UC
Stickleback    GCCUUUCCGAUGGCAGGUGUAGCUACAGAGAAACCUGGUCA---C-----UC
Medaka         ACUUUUUGGAUGGCAGGUGUAGCUACAGAGAAGCCUGUUUCCAUC-----UC
Zebrafish      GCACU---GAUAGCGGGCGUAGAUACAGAGAAGCUUUGCUCCCUC-----AC
fold           (((((((((..(((((((((....))......)))))))....)))))))))
pair symbols   abcdefghi  klmnopqrs    sr      qponmlk    ihgfedcba
```

No change

⬛ Conserved paired nucleotide

▨ Conserved unpaired nucleotide

Changes characteristic of RNA evolution

🟦 Silent G•U substitution

🟪 Silent substitution in unpaired base

🟩 Silent base-preserving double substitution

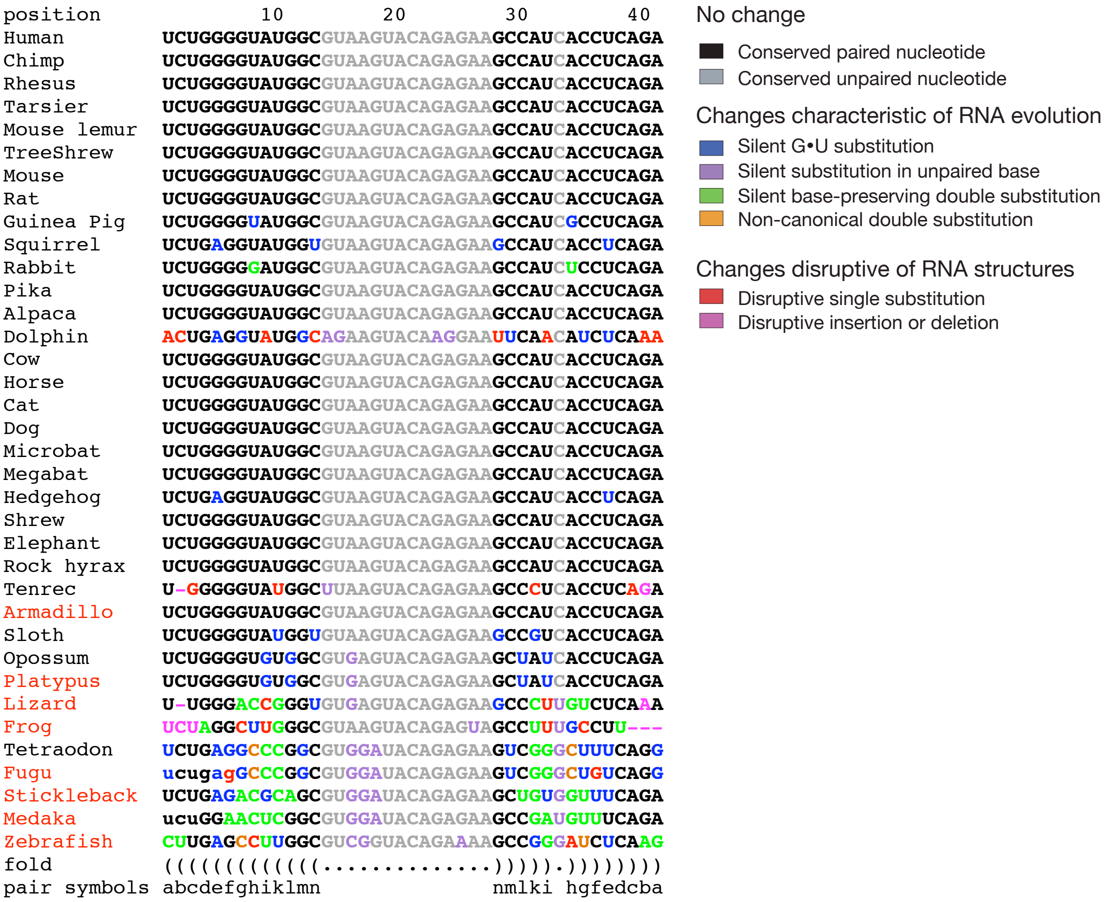
🟧 Non-canonical double substitution

Changes disruptive of RNA structures

🟥 Disruptive single substitution

▩ Disruptive insertion or deletion

**Figure S11: Substitution evidence for *MAT2A* 3′UTR family, hairpin C.**
Segment of 41-way genomic alignment color-coded according substitution evidence for structure prediction. Species not used for inference by EvoFold are shown in red.



```
position                10        20        30        40
Human        UCUGGGGUAUGGCGUAAGUACAGAGAAGCCAUCACCUCAGA
Chimp        UCUGGGGUAUGGCGUAAGUACAGAGAAGCCAUCACCUCAGA
Rhesus       UCUGGGGUAUGGCGUAAGUACAGAGAAGCCAUCACCUCAGA
Tarsier      UCUGGGGUAUGGCGUAAGUACAGAGAAGCCAUCACCUCAGA
Mouse lemur  UCUGGGGUAUGGCGUAAGUACAGAGAAGCCAUCACCUCAGA
TreeShrew    UCUGGGGUAUGGCGUAAGUACAGAGAAGCCAUCACCUCAGA
Mouse        UCUGGGGUAUGGCGUAAGUACAGAGAAGCCAUCACCUCAGA
Rat          UCUGGGGUAUGGCGUAAGUACAGAGAAGCCAUCACCUCAGA
Guinea Pig   UCUGGGGUAUGGCGUAAGUACAGAGAAGCCAUCGCCUCAGA
Squirrel     UCUGAGGUAUGGUGUAAGUACAGAGAAGCCAUCACCUCAGA
Rabbit       UCUGGGGGAUGGCGUAAGUACAGAGAAGCCAUCUCCUCAGA
Pika         UCUGGGGUAUGGCGUAAGUACAGAGAAGCCAUCACCUCAGA
Alpaca       UCUGGGGUAUGGCGUAAGUACAGAGAAGCCAUCACCUCAGA
Dolphin      ACUGAGGUAUGGCAGAAGUACAAGGAAUUCAACAUCUCAAA
Cow          UCUGGGGUAUGGCGUAAGUACAGAGAAGCCAUCACCUCAGA
Horse        UCUGGGGUAUGGCGUAAGUACAGAGAAGCCAUCACCUCAGA
Cat          UCUGGGGUAUGGCGUAAGUACAGAGAAGCCAUCACCUCAGA
Dog          UCUGGGGUAUGGCGUAAGUACAGAGAAGCCAUCACCUCAGA
Microbat     UCUGGGGUAUGGCGUAAGUACAGAGAAGCCAUCACCUCAGA
Megabat      UCUGGGGUAUGGCGUAAGUACAGAGAAGCCAUCACCUCAGA
Hedgehog     UCUGAGGUAUGGCGUAAGUACAGAGAAGCCAUCACCUCAGA
Shrew        UCUGGGGUAUGGCGUAAGUACAGAGAAGCCAUCACCUCAGA
Elephant     UCUGGGGUAUGGCGUAAGUACAGAGAAGCCAUCACCUCAGA
Rock hyrax   UCUGGGGUAUGGCGUAAGUACAGAGAAGCCAUCACCUCAGA
Tenrec       U-GGGGGUAUGGCUUAAGUACAGAGAAGCCCUCACCUCAGA
Armadillo    UCUGGGGUAUGGCGUAAGUACAGAGAAGCCAUCACCUCAGA
Sloth        UCUGGGGUAUGGUGUAAGUACAGAGAAGCCGUCACCUCAGA
Opossum      UCUGGGGUGUGGCGUGAGUACAGAGAAGCUAUCACCUCAGA
Platypus     UCUGGGGUGUGGCGUGAGUACAGAGAAGCUAUCACCUCAGA
Lizard       U-UGGGACCGGGUGUGAGUACAGAGAAGCCCUUGUCUCAAA
Frog         UCUAGGCUUGGGCGUAAGUACAGAGUAGCCUUUGCCUU---
Tetraodon    UCUGAGGCCCGGCGUGGAUACAGAGAAGUCGGGCUUUCAGG
Fugu         ucugagGCCCGGCGUGGAUACAGAGAAGUCGGGCUGUCAGG
Stickleback  UCUGAGACGCAGCGUGGAUACAGAGAAGCUGUGGUUUCAGA
Medaka       ucuGGAACUCGGCGUGGAUACAGAGAAGCCGAUGUUUCAGA
Zebrafish    CUUGAGCCUUGGCGUCGGUACAGAAAAGCCGGGAUCUCAAG
fold         (((((((((((((..............))))).))))))))
pair symbols abcdefghiklmn              nmlki hgfedcba
```

No change

- ■ Conserved paired nucleotide
- ■ Conserved unpaired nucleotide

Changes characteristic of RNA evolution

- 🟦 Silent G•U substitution
- 🟪 Silent substitution in unpaired base
- 🟩 Silent base-preserving double substitution
- 🟧 Non-canonical double substitution

Changes disruptive of RNA structures

- 🟥 Disruptive single substitution
- 🟪 Disruptive insertion or deletion

**Figure S12: Substitution evidence for *MAT2A* 3′UTR family, hairpin D.**
Segment of 41-way genomic alignment color-coded according substitution evidence for structure prediction. Species not used for inference by EvoFold are shown in red.

```
position               10        20        30        40        50        60
Human         CCAGCAUUCCCAGGUAGGCCAAGGUGUCCUACAGAAAAACCUUGGGUUAGACCUACAGGGGGUCUGG
Chimp         CCAGCAUUCCCAGGUAGGCCAAGGUGUCCUACAGAAAAACCUUGGGUUAGACCUACAGGGGGUCUGG
Rhesus        CCAGCAUUCCCAGGUAGGCCAAGGUGUCCUACAGAAAAACCUUGGGUUAGACCUACAGGGGGUCUGG
Tarsier       CCAGCAUUCCCAGGUAAGCCAAGGUGCCCUACAGAAAAACCUUGGGUUAGACCUAGAGGGGGUCUGA
Mouse lemur   CCAGCAUUCCCAGGUAGGCCAAGGUGUCCUACAGAAAAACCUUGGGUUAGACCUACAGGGGGUCUGG
TreeShrew     CCAGCAUUCCCAGGUAGGCCAAGGUGUCCUACAGAAAAACCUUGGGUUAGUCCUACAGGGGGUCUGG
Mouse         CCAGCAUUCCCAGGUAGGCCAAGGUGUCCUACAGAAAAACCUUGGGUUAGACCUACAGGGGGUCUGG
Rat           CCAGCAUUCCCAGGUAGGCCAAGGUGUCCUACAGAAAAACCUUGGGUUAGACCUACAGGGGGUCUGG
Guinea Pig    CCAGCAUUCCCAGGUAGGCCAAGGUGUCCUACAGAAAAACCUUGGGUUAGACCUACAGGGGGUCUGG
Squirrel      CCAGCAUUCCCAGGUAGGCCAAGGUGUCCUACAGAAAAACCUUGGGUUAGACCUACAGGGGGUCUGG
Rabbit        CCAGCAUUCCCAGGUAGGCCAAGGUGUCCUACAGAAAAACCUUGGGUUAGACCUACAGGGGGUCUGG
Pika          CCAGCAUUCCCAGGUAGGCCAAGGUGUCCUACAGAAAAACCUUGGGUUGGACCUACAGGGGGUCUGG
Alpaca        CCAGCAUUCCCAGGUAGGCCAAGGUGUCCUACAGAAAAACCUUGGGUUAGACCUACAGGGGGUCUGG
Dolphin       UCAGCAUUCAUGAAAGGUCAAGGCAUCCUACGAAAAAACCUUGAGUGAGAUCUAUAGGCAGUCUGG
Cow           CCAGCAUUCCCAGGUAGGCCAAGGUGUCCUACAGAAAAACCUUGGGUUAGACCUACAGGGGUUCUGG
Horse         CCAGCAUUCCCAGGUAGGCCAAGGUGUCCUACAGAAAAACCUUGGGUUAGACCUACGGGGGGUCUGG
Cat           CCAGCAUUCCCAGGUAGGCCAAGGUGUCCUACAGAAAAACCUUGGGUUAGACCUACAGGGGGUCUGG
Dog           CCAGCAUUCCCAGGUAGGCCAAGGUGUCCUACAGAAAAACCUUGGGUUAGACCUACAGGGGGUCUGG
Microbat      CCAGCAUUCCCAGGUAGGCCAAGGUGUCCUACAGAAAAACCUUGGGUUAGACCUACAGGGGUUCUGG
Megabat       CCAGCGUUCCCAGGUAGGCCAAGGUGUCCUACAGAAAAACCUUGGGUUAGACCUACAGGGGGUCUGG
Hedgehog      CCAGCAUUCCCAGGUAGGCCAAGGUGUCCUACAGAAAAACCUUGGGUUAGACCUACAGGGGGUCUGG
Shrew         CCAGCAUUCCCAGGUAGGCCAAGGUGUCCUACAGAAAAACCUUGGGUUAGACCUACAGGGGGUCUGG
Elephant      CCAGCAUUCCCAGGUAGGCCAAGGUGUCCUACAGAAAAACCUUGGGUUAGACCUACAGGGGGUCUGG
Rock hyrax    CCAGCAUUCCCAGGUAGGCCAAGGUGUCCUACAGAAAAACCUUGGGUUAGACCUACAGGGGGUCUGG
Tenrec        CCAGCAUUCCCAGGUAGGCCGAGGUGUCCUACAGAAAAGCCUUGGGUUAGACCUACA-GGGGUCU-G
Armadillo     CCAGCAUUCCCAGGUAGGCCAAGGUGUCCUACAGAAAAACCUUGGGUUAGACCUACAGGGGGUCUGG
Sloth         CCAGCAUUCCCAGGUAGGCCAAGGUGUCCUACAGAAAAACCUUGGGUUAGACCUACAGGGGGUCUGG
Opossum       CCAGCAUUCCCAGGUAGGCCAAGGUGUCCUACAGAAGAACCUUGGGUCACACCUACAGGGGGUCUGG
Platypus      CCAGCAUUCCCAGGUAGGCCAGGGUGUCCUACAGAAAAACCUUGGGUCGCACCUUUGGGGGUUCUGG
Lizard        CCAGCACUCCCUGGUAGGCCAAGGUGUCCUACAGAAAAGCCUUGGGUCCCAUCCGCAGGGGGUCUGG
Frog          CUAGUGCUCUCUGCCAGCUCAAGGUGUAGUACAGAAAAACCUUGGCUGUAUUCUGCAGGGGUCCUGG
Tetraodon     UCAGUGUUCCCGGAGGCCCCAGGGUGUCCUACAGAAAAGCUCUGGUCAGCCUCGGCC--------GC
Fugu          UCGGUGUUCCUGGGGGAUCCGGGGUGUUGUACAGAAAAGCUCUGGUCAACCUCGGCA--------GC
Stickleback   UCAGUGUUUCUUGAGGCGCCAGAGUGUUGUACAGAAAAGCUUGGGCCGCCUCAAC------------
Medaka        UCAGUGUUCCUUGAGGGCCCAGGGUGUUGUACAGAAAAGCUUGGUCUACCUCAA-------------
Zebrafish     AUGGUGUCCCGCGCCGGCCCUGGGUGUCCUACAG-AAAGCUUUGGGCUCUUGCCACA--------GG
fold          ((((.((((((..((((((((((((((...))......)))))))......))))).))))))))))
pair symbols  abcd efghik  lmnopqrstuvwxy   yx      wvutsrq      ponml kihgfedcba
                           ↑                                          ↑
```

No change

⬛ Conserved paired nucleotide

▨ Conserved unpaired nucleotide

Changes characteristic of RNA evolution

🟦 Silent G•U substitution

🟪 Silent substitution in unpaired base

🟩 Silent base-preserving double substitution

🟧 Non-canonical double substitution

Changes disruptive of RNA structures

🟥 Disruptive single substitution

🟪 Disruptive insertion or deletion

**Figure S13: Substitution evidence for *MAT2A* 3′UTR family, hairpin E.**
Segment of 41-way genomic alignment color-coded according substitution evidence for structure prediction. Species not used for inference are shown in red. Only human was used for inference in this case, since the structure was found by a homology search. The extension of the originally predicted structure, by single sequence energy minimization folding, is indicated with red arrows at the bottom of the alignment.

```
position             10        20        30        40        50        60
Human        CAUGGAGAAAGCUGACUUGGCUGGUGUGGUACAGAGAAGCCAGCUUGUUUACAUGCUUAUUCCAUG
Cow          CAUGGAGAAAGCUGACUUGGCUGGUGUGGUACAGAGAAGCCAGCUUGUUUACAUGCUUAUUCCAUG
Horse        CAUGGAGAAAGCUGACUUGGCUGGUGUGGUACAGAGAAGCCAGCUUGUUUACAUGCUUAUUCCAUG
Alpaca       CAUGGAGAAAGCUGACUUGGCUGGUGUGGUACAGAGAAGCCAGCUUGUUUACAUGCUUAUUCCAUG
Cat          CAUGGAGAAAGCUGACUUGGCUGGUGUGGUACAGAGAAGCCAGCUUGUUUACAUGCUUAUUCCAUG
Pika         CAUGGAGGAAGCGGACUUGGCUGGUGUGGUACAGAGAAGCCAGCUUGUUUACAUGCUUAUUCCAUG
Dog          CAUGGAGAAAGCUGACUUGGCUGGUGUGGUACAGAGAAGCCAGCUUGUUUACAUGCUUAUUCCAUG
Rabbit       CAUGGAGAAAGCGGACUUGGCUGGUGUGGUACAGAGAAGCCAGCUUGUUUACAUGCUUCAUCCAUG
Microbat     CAUGGAGAAAGCUGACUUGGCUGGUGUGGUACAGAGAAGCCAGCUUGUUUACAUGCUUAUUCCAUG
Squirrel     CGUGGAGAGAGCGGACUUGGCUGGCGUGGUACAGAGAAGCCAGCUUGUUUACAUGC-UACUCCAUG
Megabat      CAUGGAAAAAGCAGACCUGGCUGGUGUGGUACAGAGAAGCCAGCUUGUUUACAUGCUUAUUUCAUG
Guinea Pig   CAUGGAGAAAGCUGACUUGGCUGGUGUGGUACAGAGAAGCCAGCUUGUUUACAUGCUUAUUCCAUG
Hedgehog     CAUGGAGAAAGCUGACUUGGCUGGUGUGGUACAGAGAAGCCAGCUUGUUUACAUGCUUGCUCCAUG
Rat          CAUGGAGAAAGCUGACUUGGCUGGUGUGGUACAGAGAAGCCAGCUUGUUUACAUGCUUAUUCCAUG
Shrew        CAUGGAGAAAGCGGACUUGGCUGGUGUGGUACAGAGAAGCCAGCUUGUUUACAUGCUUAUUCCAUG
Mouse        CAUGGAGAAAGCUGACUUGGCUGGUGUGGUACAGAGAAGCCAGCUUGUUUACAUGCUUAUUCCAUG
Elephant     CAUGGAGAAAGCUGACUUGGCUGGUGUGGUACAGAGAAGCCAGCUUGUUUACAUGCUUAUUCCAUG
TreeShrew    CAUGGAGAAAGCUGACUUGGCUGGUGUGGUACAGAGAAGCCAGCUUGUUUACAUGCUUAUUCCAUG
Rock hyrax   CAUGGAGAAAGCUGACUUGGCUGGUGUGGUACAGAGAAGCCAGCUUGUUUACAUGCUUAUUCCAUG
Mouse lemur  CAUGGAGAAAGCUGACUUGGCUGGUGUGGUACAGAGAAGCCAGCUUGUUUACAUGCUUAUUCCAUG
Tenrec       CAUGGAG-AGGCUGACUUGGCCGGUGUGCUACAGAGAAGCCAGCUUGUUGACAUGCUCCUCCAUUG
Tarsier      CAUGGAGAAAGCUGACUUGGCUGGUGUGGUACAGAAAAGCCAGCUUGUUUACAUGCUUAUUCCAUG
Armadillo    CAUGGAGAAAGCUGACUUGGCUGGUGUGGUACAGAGAAGCCAGCUUGUUUACAUGCUUAUUCCAUG
Opossum      CAUGGAGGAAGCUGACUUGGCUGGUGUGGUACAGAGAAGCCAGCUUGUUUACAUGCCUACUCCAUG
Rhesus       CAUGGAGAAAGCUGACUUGGCUGGUGUGGUACAGAGAAGCCAGCUUGUUUACAUGCUUAUUCCAUG
Platypus     CGUGGGGAGAGCUGACUUGGCUGGUGUUGUACAGAGAAGCCAGCUUGUUUACAGGCUCAUCCCAUG
Lizard       CCUGGGAGGGGCUGACUUGGCCGGUGUGGUACAGAGAAGCCGGCUUGUUUACUUACCCGUCCCACG
Frog         ....ACAGAAGCUGGCUUUACCAGUGUUGUACAGAGAAACUGG-GCGUUUACAAGCUU--UCCUUG
Chimp        CAUGGAGAAAGCUGACUUGGCUGGUGUGGUACAGAGAAGCCAGCUUGUUUACAUGCUUAUUCCAUG
Tetraodon    CCCCGGCAA--C--AUUUGGUUGGUGUAGUACAGAGAAACCAGCGUGUUUACACGACCAUUCCCUG
Fugu         aauuGGCAA--C--AUUUGGUUGGUGUAGUACAGAGAAACCAGCGUGUUUACAUGACCAUCUCGUG
Dolphin      UG---------------UGGCUAGUGUGGUACA.................................
Stickleback  ....----------AUCCGGUUGGUGUUGUACAGAGAAACCAGCUUGUUUACGCAGCCAUGUCAU-
Zebrafish    CACAGGAAG--CGAGUCUAGCUGGUGUUGUACAGAGAAACUGGCUCGUUUACAUGUCUGUCUGCUG
Medaka       ....----------ACGUGGCUGGUGUUGUACAGAGAAACUAGCUUGUUUACAAAGCCAUCCCGCA
fold         (((((((.((((.(((..(((((((((...))......))))))).))).....)))).)))))))
pair symbols abcdefg hikl mno  pqrstuvwx   xw      vutsrqp onm     lkih gfedcba
                                 ↑                    ↑
```

### No change

⬛ Conserved paired nucleotide

⬜ Conserved unpaired nucleotide

### Changes characteristic of RNA evolution

🟦 Silent G•U substitution

🟪 Silent substitution in unpaired base

🟩 Silent base-preserving double substitution

🟧 Non-canonical double substitution

### Changes disruptive of RNA structures

🟥 Disruptive single substitution

🟪 Disruptive insertion or deletion

**Figure S14: Substitution evidence for *MAT2A* 3′UTR family, hairpin F.**
Segment of 41-way genomic alignment color-coded according substitution evidence for structure prediction. Species not used for inference are shown in red. Only human was used for inference in this case, since the structure was found by a homology search. The extension of the originally predicted structure, by single sequence energy minimization folding, is indicated with red arrows at the bottom of the alignment.

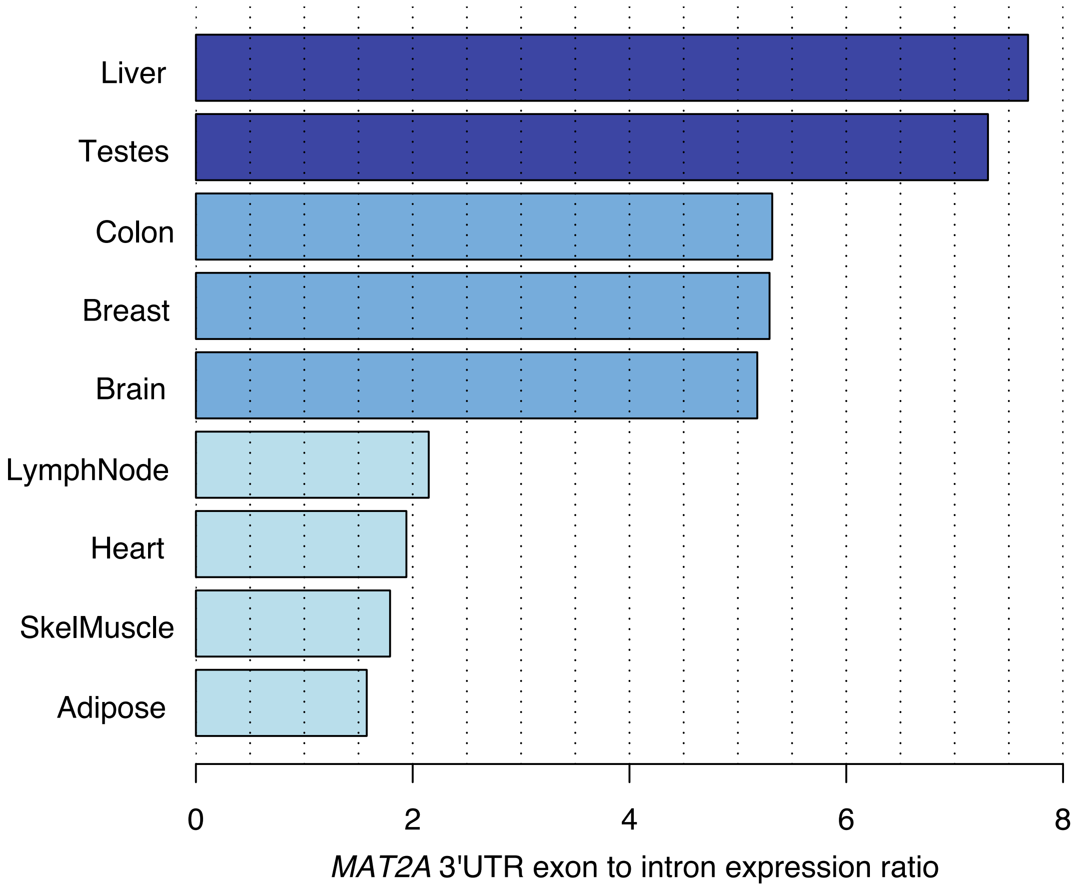


**Figure S15: Tissue-specific ratios of *MAT2A* expression comparing upstream and downstream regions of the 3′UTR.**

The upstream and downstream regions were defined as chr2:85624360-85625384 and chr2: 85625385-85625913, respectively. The downstream region overlies hairpins C-F; the upstream region overlies a putative alternative intron in the 3′UTR, as defined by EST and spliced RNA-seq evidence (see figure 1). RNA-seq data from (Wang et al. 2008).

Overall tissue specificity (p-value=8e-14; chi-squared test) is indicative of a functional role for *MAT2A* 3′UTR isoforms. The rank ordering of expression ratio broadly matches the known tissue distribution of methionine adenosyltransferase (Eloranta 1977) (across 6 tissues: liver and testes the highest ranking, with skeletal muscle the lowest ranking). Spearman correlation coefficient was 0.94, p-value=8e-3. Note, than an analysis limited to upstream and downstream regions each 350 bp long flanking the 3′ splice site, which excludes the most 5′ known alternative polyadenylation site (upstream region:chr2:85625034-85625384; downstream region:chr2:85625385-85625735), similarly showed Spearman correlation coefficient of 0.89, p-value=0.02.

```
1    CGCCCGCCUG CUACGAGUAG AACGCUGUCC GCAGCUUGCG CAUUUCGCAG CCGCUGCCGC
61   CUCGCCGCUG CUCCUUCGUA AGGCCACUUC CGCACACCGA CACCAACAUG AACGGACAGC
121  UCAACGGCUU CCACGAGGCG UUCAUCGAGG AGGGCACAUU CCUUUUCACC UCAGAGUCGG
181  UCGGGGAAGG CCACCCAGAU AAGAUUUGUG ACCAAAUCAG UGAUGCUGUC CUUGAUGCCC
241  ACCUUCAGCA GGAUCCUGAU GCCAAAGUAG CUUGUGAAAC UGUUGCUAAA ACUGGAAUGA
301  UCCUUCUUGC UGGGGAAAUU ACAUCCAGAG CUGCUGUUGA CUACCAGAAA GUGGUUCGUG
361  AAGCUGUUAA ACACAUUGGA UAUGAUGAUU CUUCCAAAGG UUUUGACUAC AAGACUUGUA
421  ACGUGCUGGU AGCCUUGGAG CAACAGUCAC CAGAUAUUGC UCAAGGUGUU CAUCUUGACA
481  GAAAUGAAGA AGACAUUGGU GCUGGAGACC AGGGCUUAAU GUUUGGCUAU GCCACUGAUG
541  AAACUGAGGA GUGUAUGCCU UUAACCAUUG UCUUGGCACA CAAGCUAAAU GCCAAACUGG
601  CAGAACUACG CCGUAAUGGC ACUUUGCCUU GGUUACGCCC UGAUUCUAAA ACUCAAGUUA
661  CUGUGCAGUA UAUGCAGGAU CGAGGUGCUG UGCUUCCCAU CAGAGUCCAC ACAAUUGUUA
721  UAUCUGUUCA GCAUGAUGAA GAGGUUUGUC UUGAUGAAAU GAGGGAUGCC CUAAAGGAGA
781  AAGUCAUCAA AGCAGUUGUG CCUGCGAAAU ACCUUGAUGA GGAUACAAUC UACCACCUAC
841  AGCCAAGUGG CAGAUUUGUU AUUGGUGGGC CUCAGGGUGA UGCUGGUUUG ACUGGACGCA
901  AAAUCAUUGU GGACACUUAU GGCGGUUGGG GUGCUCAUGG AGGAGGUGCC UUUUCAGGAA
961  AGGAUUAUAC CAAGGUCGAC CGUUCAGCUG CUUAUGCUGC UCGUUGGGUG GCAAAAUCCC
1021 UUGUUAAAGG AGGUCUGUGC CGGAGGGUUC UUGUUCAGGU CUCUUAUGCU AUUGGAGUUU
1081 CUCAUCCAUU AUCUAUCUCC AUUUUCCAUU AUGGUACCUC UCAGAAGAGU GAGAGAGAGC
1141 UAUUAGAGAU UGUGAAGAAG AAUUUCGAUC UCCGCCCUGG GGUCAUUGUC AGGGAUCUGG
1201 AUCUGAAGAA GCCAAUUUAU CAGAGGACUG CAGCCUAUGG CCACUUUGGU AGGGACAGCU
1261 UCCCAUGGGA AGUGCCCAAA AAGCUUAAAU AUUGAAAGUG UUAGCCUUUU UUCCCCAGAC
1321 UUGUUGGCGU AGGCUACAGA GAAGCCUUCA AGCUCUGAGG GAAAGGGCCC UCCUUCCUAA
1381 AUUUUCCUGU CCUCUUUCAG CUCCUGACCA GUUGCAGUCA CUCUAGUCAA UGACAUGAAU
1441 UUUAGCUUUU GUGGGGGACU GUAAGUUGGG CUUGCUAUUC UGUCCCUAGG UGUUUUGUUC
1501 ACCAUUAUAA UGAAUUUAGU GAGCAUAGGU GAUCCAUGUA ACUGCCUAGA AACAACACUG
1561 UAGUAAAUAA UGCUUUGAAA UUGAACCUUU GUGCCCUAUC ACCCAACGCU CCAAAGUCAU
1621 AAUUGCAUUG ACUUUCCCCA CCAGAUGCUG AAAAUGUCCU UGUGAUGUGC ACGUAAAGUA
1681 CUUGUAGUUC CACUUAUAGC CUCUGUCUGG CAAUGCCACA GCCCUGUCAG CAUGAAUUUG
1741 UAAUGUCUUG AGCUCUAUUA UGAAUGUGAA GCCUUCCCCU UAUCCUCCCU GUAACUUGAU
1801 CCAUUUCUAA UUAUGUAGCU CUUUGUCAGG GAGUGUUCCC UAUCCAAUCA AUCUUGCAUG
1861 UAACGCAAGU UCCCAGUUGG AGCUCCAGCC UGACAUCAAA AAAGGCAGUU ACCAUUAAAC
1921 CAUCUCCCUG GUGCUUAUGC UCUUAAUUGC CACCUCUAAC AGCACCAAAU CAAAAUCUCU
1981 CCACUUUCAG CUGUCUUUUG GAGGACGUAC GUAAUAAGGU UUUAAUUUAG UAAACCAAUC
2041 CUAUGCAUGG UUUCAGCACU AGCCAAACCU CACCAACUCC UAGUUCUAGA AAAACAGGCA
2101 CUUGGCAGCC UUGUGAUGUC AUACAGAGAA GUCACAGGGC AGUACCUGAG GGUCUGUAGG
2161 UUGCACACUU UGGUACCAGA UAACUUUUUU UUUUCUUUAU AAGAAAGCCU GAGUACUCCA
2221 CACUGCACAA UAACUCCUCC CAGGGUUUUA ACUUUGUUUU AUUUUCAAAA CCAGGUCCAA
2281 UGAGCUUUCU GAACAGCUGG UGUAGCUACA GAGAAACCAG CUUCCUUCAG AGAGCAGUGC
2341 UUUUGGCGGG GAGGAGGAAA UCCCUUCAUA CUUGAACGUU UUCUAAUUGC UUAUUUAUUG
2401 UAUUCUGGGG UAUGGCGUAA GUACAGAGAA GCCAUCACCU CAGAUGGCAG CUUUUAAAAG
2461 AUUUUUUUUU UUUCUCUCAA CACCAUGAUU CCUUUAACAA CAUGUUUCCA GCAUUCCCAG
2521 GUAGGCCAAG GUGUCCUACA GAAAAACCUU GGGUUAGACC UACAGGGGGU CUGGCUGGUG
2581 UUAACAGAAG GGAGGGCAGA GCUGGUGCGG CUGGCCAUGG AGAAAGCUGA CUUGGCUGGU
2641 GUGGUACAGA GAAGCCAGCU UGUUUACAUG CUUAUUCCAU GACUGCUUGC CCUAAGCAGA
2701 AAGUGCCUUU CAGGAUCUAU UUUUGGAGGU UUAUUACGUA UGUCUGGUUC UCAAUUCCAA
2761 CAGUUUAAUG AAGAUCUAAA UAAAAUGCUA GGUUCUACCU UAAAAAAAAA AAAAAAAAA
```

```
Predicted stem of hairpin
Predicted loop of hairpin
Start codon corresponding to methionine adenosyltransferase II, alpha
Stop codon corresponding to methionine adenosyltransferase II, alpha
```

186 *MAT2A* (contains RNA hairpin A) 186 nt

```
GGGACAGCUUCCCAUGGGAAGUGCCCAAAAAGCUUAAAUAUUGAAAGUGUUAGCCUUUUUUCCCCAGACUUGUUGGCGUAGGCUACAGAGAAGCCUUC
AAGCUCUGAGGGAAAGGGCCCUCCUUCCUAAAUUUUCCUGUCCUCUUUCAGCUCCUGACCAGUUGCAGUCACUCUAGUCAAUGACAUG
```

411 *MAT2A* (contains RNA hairpins C, D, E) 411 nt

```
GGCCUGAGUACUCCACACUGCACAAUAACUCCUCCCAGGGUUUUAACUUUGUUUUAUUUUCAAAACCAGGUCCAAUGAGCUUUCUGAACAGCUGGUGU
AGCUACAGAGAAACCAGCUUCCUUCAGAGAGCAGUGCUUUUGGCGGGGAGGAGGAAAUCCCUUCAUACUUGAACGUUUUCUAAUUGCUUAUUUAUUGU
AUUCUGGGGUAUGGCGUAAGUACAGAGAAGCCAUCACCUCAGAUGGCAGCUUUUAAAAGAUUUUUUUUUUUUCUCUCAACACCAUGAUUCCUUUAACA
ACAUGUUUCCAGCAUUCCCAGGUAGGCCAAGGUGUCCUACAGAAAAACCUUGGGUUAGACCUACAGGGGGUCUGGCUGGUGUUAACAGAAGGGAGGGC
AGAGCUGGUGCGGCUGGCC
```

358 *MAT2A* (contains RNA hairpins D, E, F) 358 nt

```
GGGAGGAGGAAAUCCCUUCAUACUUGAACGUUUUCUAAUUGCUUAUUUAUUGUAUUCUGGGGUAUGGCGUAAGUACAGAGAAGCCAUCACCUCAGAUG
GCAGCUUUUAAAAGAUUUUUUUUUUUUCUCUCAACACCAUGAUUCCUUUAACAACAUGUUUCCAGCAUUCCCAGGUAGGCCAAGGUGUCCUACAGAAA
AACCUUGGGUUAGACCUACAGGGGGUCUGGCUGGUGUUAACAGAAGGGAGGGCAGAGCUGGUGCGGCUGGCCAUGGAGAAAGCUGACUUGGCUGGUGU
GGUACAGAGAAGCCAGCUUGUUUACAUGCUUAUUCCAUGACUGCUUGCCCUAAGCAGAAAGUGC
```

268 *MAT2A* (contains RNA hairpins D, E) 268 nt

```
GGGAGGAGGAAAUCCCUUCAUACUUGAACGUUUUCUAAUUGCUUAUUUAUUGUAUUCUGGGGUAUGGCGUAAGUACAGAGAAGCCAUCACCUCAGAUG
GCAGCUUUUAAAAGAUUUUUUUUUUUUCUCUCAACACCAUGAUUCCUUUAACAACAUGUUUCCAGCAUUCCCAGGUAGGCCAAGGUGUCCUACAGAAA
AACCUUGGGUUAGACCUACAGGGGGUCUGGCUGGUGUUAACAGAAGGGAGGGCAGAGCUGGUGCGGCUGGCC
```

**Figure S16: RNA constructs used for in-line probing analyses (5′ to 3′).**
*Homo sapiens* methionine adenosyltransferase II, alpha, mRNA (cDNA clone MGC:2907 IMAGE:3010820)

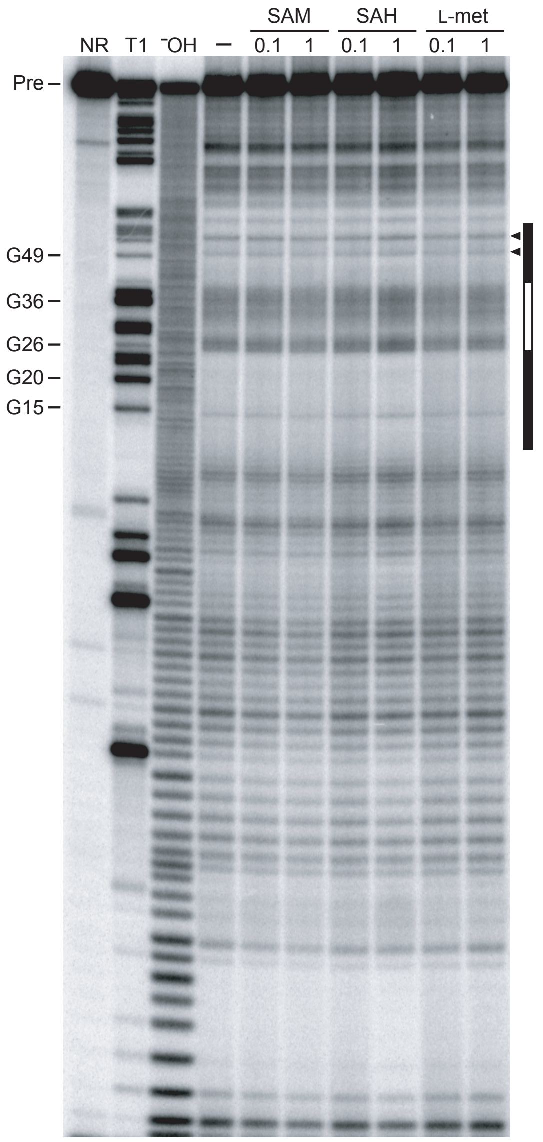


**Figure S17: In-line probing analysis of 186 *MAT2A*.**
RNA cleavage products resulting from spontaneous transesterification during incubations in the absence (-) of any candidate ligand or in the presence of SAM, *S*-adenosylhomocysteine (SAH), and L-methionine (L-met), each tested at concentrations of 0.1 mM and 1 mM, were resolved by denaturing 10% PAGE. NR, no reaction; T1, partial digest with RNase T1; ¯OH, partial alkaline digest; Pre, precursor RNA. Selected bands in the T1 lane are labeled with the positions of the respective 3′ terminal guanosyl residues, according to the numbering used for hairpin A in figure 2C. Filled bars correspond to positions within hairpin A that are predicted to be largely base-paired, while the open bar corresponds to positions within the putative loop sequence. Arrowheads correspond to putative bulged nucleotides C50 and A55. This supplementary figure contains the complete autoradiogram corresponding to the analysis of 186 *MAT2A*, only part of which is depicted in figure 2D.

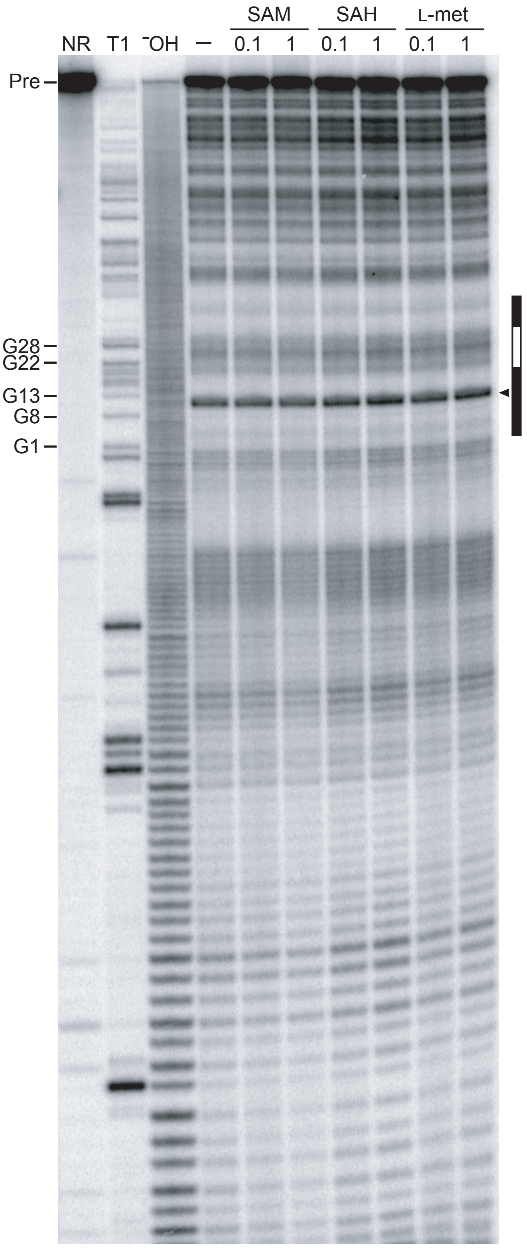


**Figure S18: In-line probing analysis of 411 *MAT2A*.**

RNA cleavage products resulting from spontaneous transesterification during incubations in the absence (-) of any candidate ligand or in the presence of SAM, *S*-adenosylhomocysteine (SAH), and L-methionine (L-met), each tested at concentrations of 0.1 mM and 1 mM, were resolved by denaturing 10% PAGE. NR, no reaction; T1, partial digest with RNase T1; ⁻OH, partial alkaline digest; Pre, precursor RNA. Selected bands in the T1 lane are labeled with the positions of the respective 3′ terminal guanosyl residues, according to the numbering used for hairpin C in figure 2C. Filled bars correspond to positions within hairpin C that are predicted to be largely base-paired, while the open bar corresponds to positions within the putative loop sequence. Arrowhead corresponds to the putative bulged nucleotide C11. Note that only the hairpin nearest the 5′ end is indicated, as the remaining hairpins within the construct cannot be mapped with sufficient resolution. The DNA template corresponding to 411 *MAT2A* was PCR-amplified from human genomic DNA using the oligodeoxynucleotide primers 5′C (5′-TAATACGACTCACTATAGGCCTGAGTACTCCACACTGCACAATAACTCC) and 3′E (5′-GGCCAGCCGCACCAGCTCTGCCCTCCCTTC).

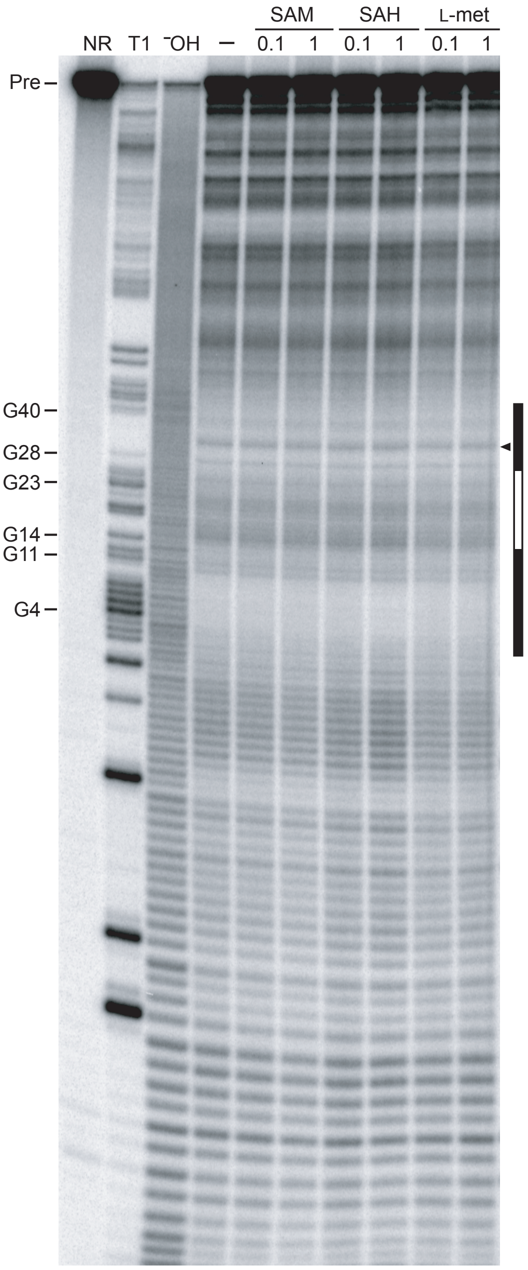


**Figure S19: In-line probing analysis of 358 *MAT2A*.**
RNA cleavage products resulting from spontaneous transesterification during incubations in the absence (-) of any candidate ligand or in the presence of SAM, *S*-adenosylhomocysteine (SAH), and L-methionine (L-met), each tested at concentrations of 0.1 mM and 1 mM, were resolved by denaturing 10% PAGE. NR, no reaction; T1, partial digest with RNase T1; ⁻OH, partial alkaline digest; Pre, precursor RNA. Selected bands in the T1 lane are labeled with the positions of the respective 3′ terminal guanosyl residues, according to the numbering used for hairpin D in figure 2C. Filled bars correspond to positions within hairpin D that are predicted to be largely base-paired, while the open bar corresponds to positions within the putative loop sequence. Arrowhead corresponds to the putative

bulged nucleotide C33. Note that only the hairpin nearest the 5′ end is indicated, as the remaining hairpins within the construct cannot be mapped with sufficient resolution. The DNA template corresponding to 358 *MAT2A* was PCR-amplified from human genomic DNA using the oligodeoxynucleotide primers 5′D (5′-TAATACGACTCACTATAGGGAGGAGGAAATCCCTTCATACTTGAACG) and 3′F (5′-GCACTTTCTGCTTAGGGCAAGCAGTCATGG).

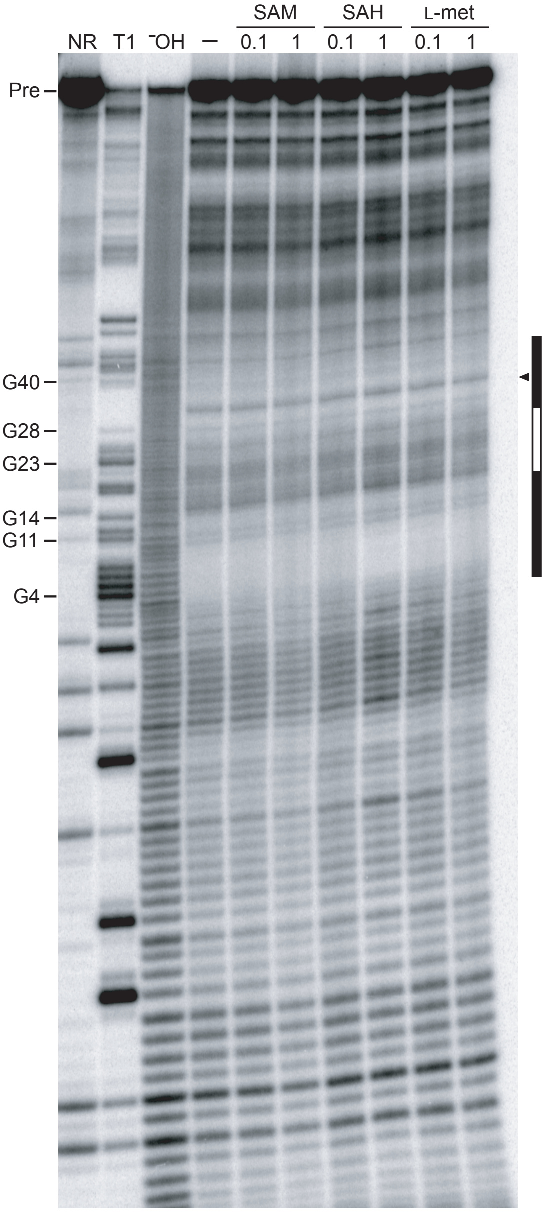
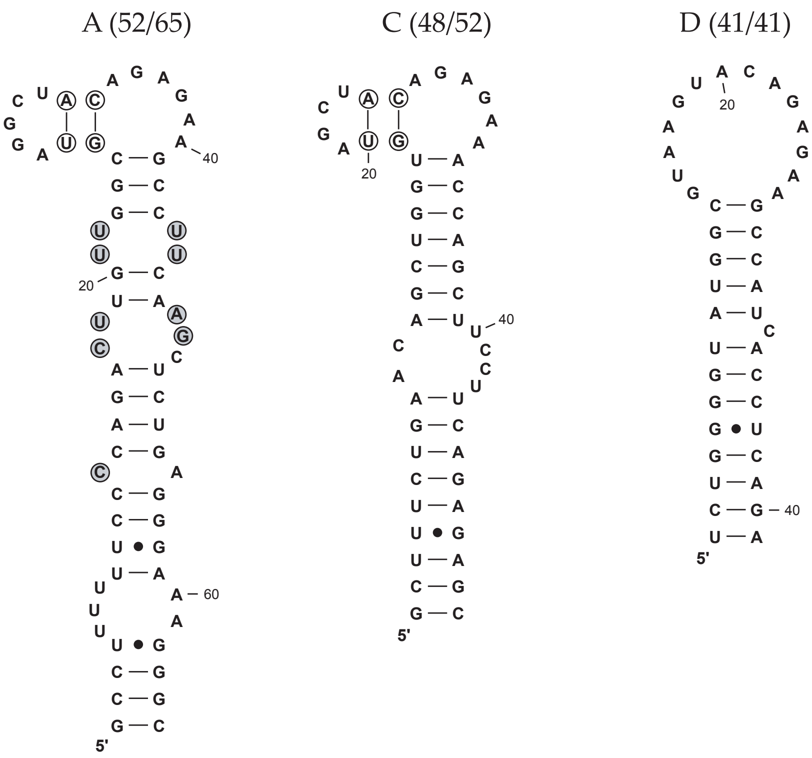


**Figure S20: In-line probing analysis of 268 *MAT2A*.**
RNA cleavage products resulting from spontaneous transesterification during incubations in the absence (-) of any candidate ligand or in the presence of SAM, *S*-adenosylhomocysteine (SAH), and L-methionine (L-met), each tested at concentrations of 0.1 mM and 1 mM, were resolved by denaturing 10% PAGE. NR, no reaction; T1, partial digest with RNase T1; ⁻OH, partial alkaline digest; Pre, precursor RNA. Selected bands in the T1 lane are labeled with the positions of the respective 3′

terminal guanosyl residues, according to the numbering used for hairpin D in figure 2C. Filled bars correspond to positions within hairpin D that are predicted to be largely base-paired, while the open bar corresponds to positions within the putative loop sequence. Arrowhead corresponds to the putative bulged nucleotide C33. Note that only the hairpin nearest the 5′ end is indicated, as the remaining hairpins within the construct cannot be mapped with sufficient resolution. The DNA template corresponding to 268 *MAT2A* was PCR-amplified from human genomic DNA using the oligodeoxynucleotide primers 5′D (5′-TAATACGACTCACTATAGGGAGGAGGAAATCCCTTCATACTTGAACG) and 3′E (5′-GGCCAGCCGCACCAGCTCTGCCCTCCCTTC).



A (52/65)

C (48/52)

D (41/41)

**Figure S21: Levels of agreement between computationally and experimentally derived *MAT2A* hairpin secondary structure predictions.**

The secondary structure models shown for *MAT2A* hairpins A, C, and D are derived computationally. Numbers in parentheses refer to the fraction of nucleotides in the hairpin whose computationally predicted Watson-Crick base-pairing status (either paired or unpaired) is supported by the structure-probing data. In cases where discrepancies may exist between the computationally predicted structure and the in-line probing analysis, the pertinent nucleotides are circled. Open circles designate residues that, although they are computationally predicted to be involved in Watson-Crick base-pairs, appear from structure-probing data to reside within relatively unstructured regions. Conversely, gray circles indicate nucleotides whose computationally predicted locations in presumably unstructured internal loops are not necessarily consistent with in-line probing results suggesting that they reside instead within relatively structured zones.

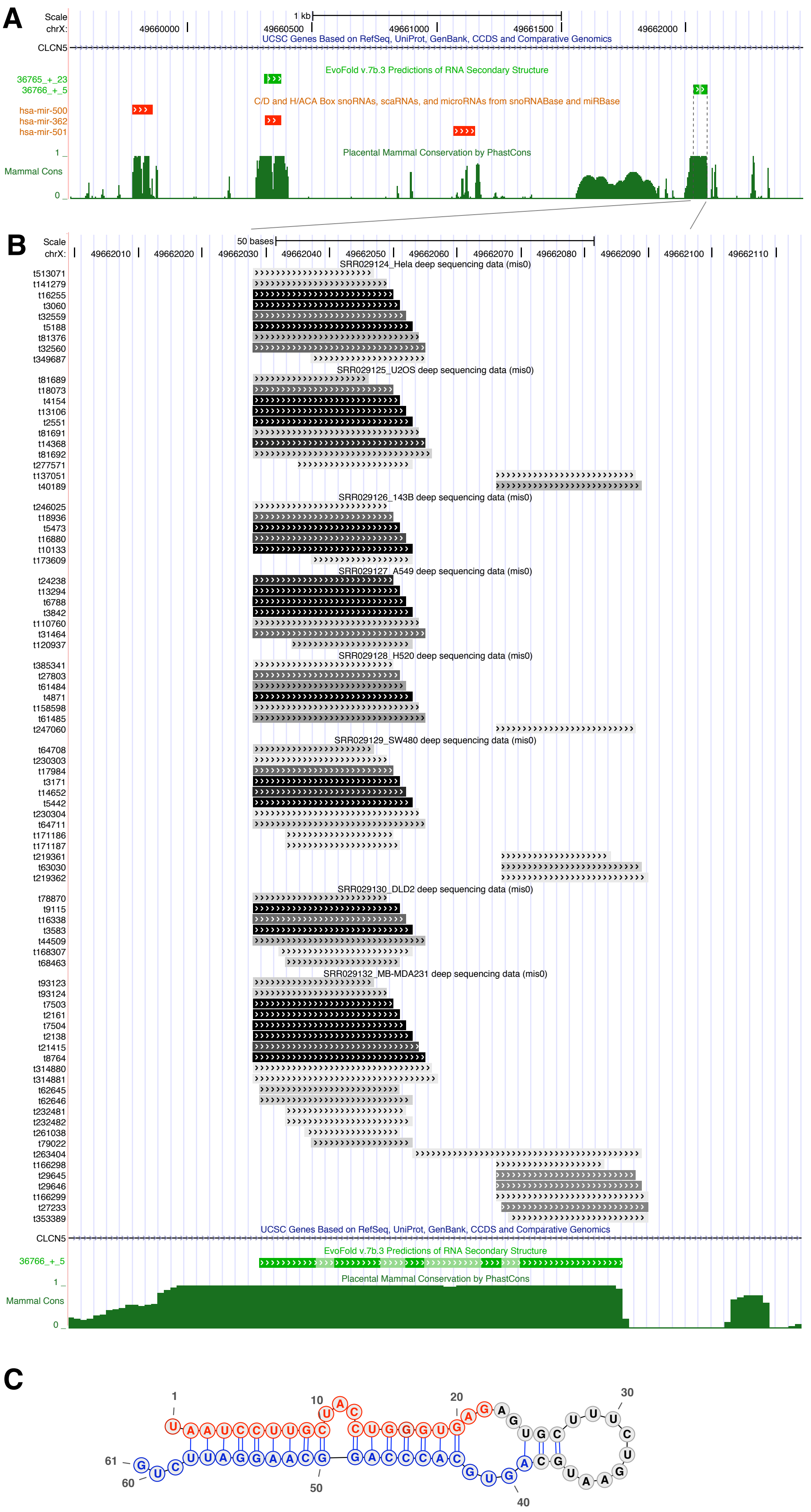

A
Scale
chrX:
1 kb
49660000
49660500
49661000
49661500
49662000
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
CLCN5
EvoFold v.7b.3 Predictions of RNA Secondary Structure
36765_+_23
36766_+_5
C/D and H/ACA Box snoRNAs, scaRNAs, and microRNAs from snoRNABase and miRBase
hsa-mir-500
hsa-mir-362
hsa-mir-501
Placental Mammal Conservation by PhastCons
Mammal Cons
1
0
B
Scale
chrX:
50 bases
49662010
49662020
49662030
49662040
49662050
49662060
49662070
49662080
49662090
49662100
49662110
SRR029124_Hela deep sequencing data (mis0)
t513071
t141279
t16255
t3060
t32559
t5188
t81376
t32560
t349687
SRR029125_U2OS deep sequencing data (mis0)
t81689
t18073
t4154
t13106
t2551
t81691
t14368
t81692
t277571
t137051
t40189
SRR029126_143B deep sequencing data (mis0)
t246025
t18936
t5473
t16880
t10133
t173609
SRR029127_A549 deep sequencing data (mis0)
t24238
t13294
t6788
t3842
t110760
t31464
t120937
SRR029128_H520 deep sequencing data (mis0)
t385341
t27803
t61484
t4871
t158598
t61485
t247060
SRR029129_SW480 deep sequencing data (mis0)
t64708
t230303
t17984
t3171
t14652
t5442
t230304
t64711
t171186
t171187
t219361
t63030
t219362
SRR029130_DLD2 deep sequencing data (mis0)
t78870
t9115
t16338
t3583
t44509
t168307
t68463
SRR029132_MB-MDA231 deep sequencing data (mis0)
t93123
t93124
t7503
t2161
t7504
t2138
t21415
t8764
t314880
t314881
t62645
t62646
t232481
t232482
t261038
t79022
t263404
t166298
t29645
t29646
t166299
t27233
t353389
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
CLCN5
EvoFold v.7b.3 Predictions of RNA Secondary Structure
36766_+_5
Placental Mammal Conservation by PhastCons
Mammal Cons
1
0
C
1
10
20
30
40
50
60
61
U A A U C C U U G C U A C C U G G G U G A G A G U G C U U U C U G A A U G C A G U G C A C C C A G G C A A G G A U U C U G

**Figure S22: RNA-seq transcript evidence for *CLCN5* putative microRNA across cell lines.**

(A) Putative microRNA genomic location, located near to the paralogous MIR362 that it clusters with, and near to MIR500 with which it shares an identical mature miR sequence.

(B) From top to bottom: HeLa human cervical carcinoma, U2OS human osteosarcoma, 143B human osteosarcoma, A549 human alveolar epithelial, H520 human non-small cell lung carcinoma, SW480 human colon adenocarcinoma, DLD2 human colon carcinoma, MB-MDA231 human breast adenocarcinoma. All RNA-seq data from (Mayr and Bartel 2009); GEO id: GSE16579.

All reads mapped using BLAT (Kent 2002) with no mismatches and multiple mapping allowed (darker shade indicates more reads). (See supplementary table S6 for details).

(C) mature miR and putative miR* shown on structure diagram. Red = mature miR, blue = miR*.

The most common miR transcript in HeLa is 23 nt and matches the mature miR-500 sequence exactly:

TAATCCTTGCTACCTGGGTGAGA

The most common miR* transcript is 23 nt:

AGTGCACCCAGGCAAGGATTCTG

The miR* sequence is unique (using UCSC Blat) which provides evidence that this locus is functional.

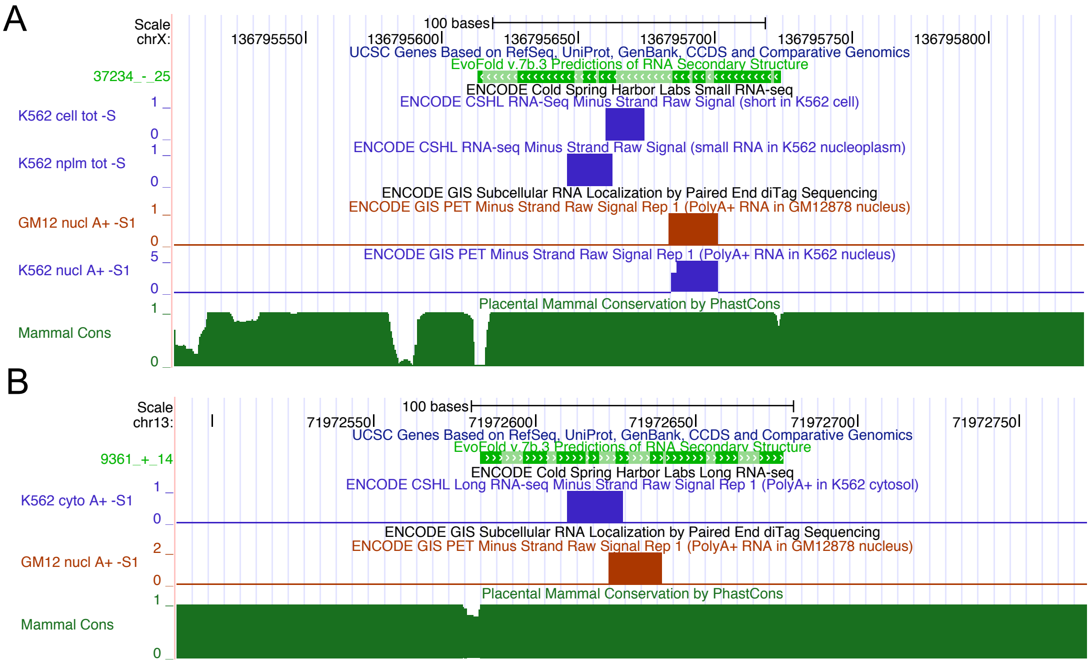
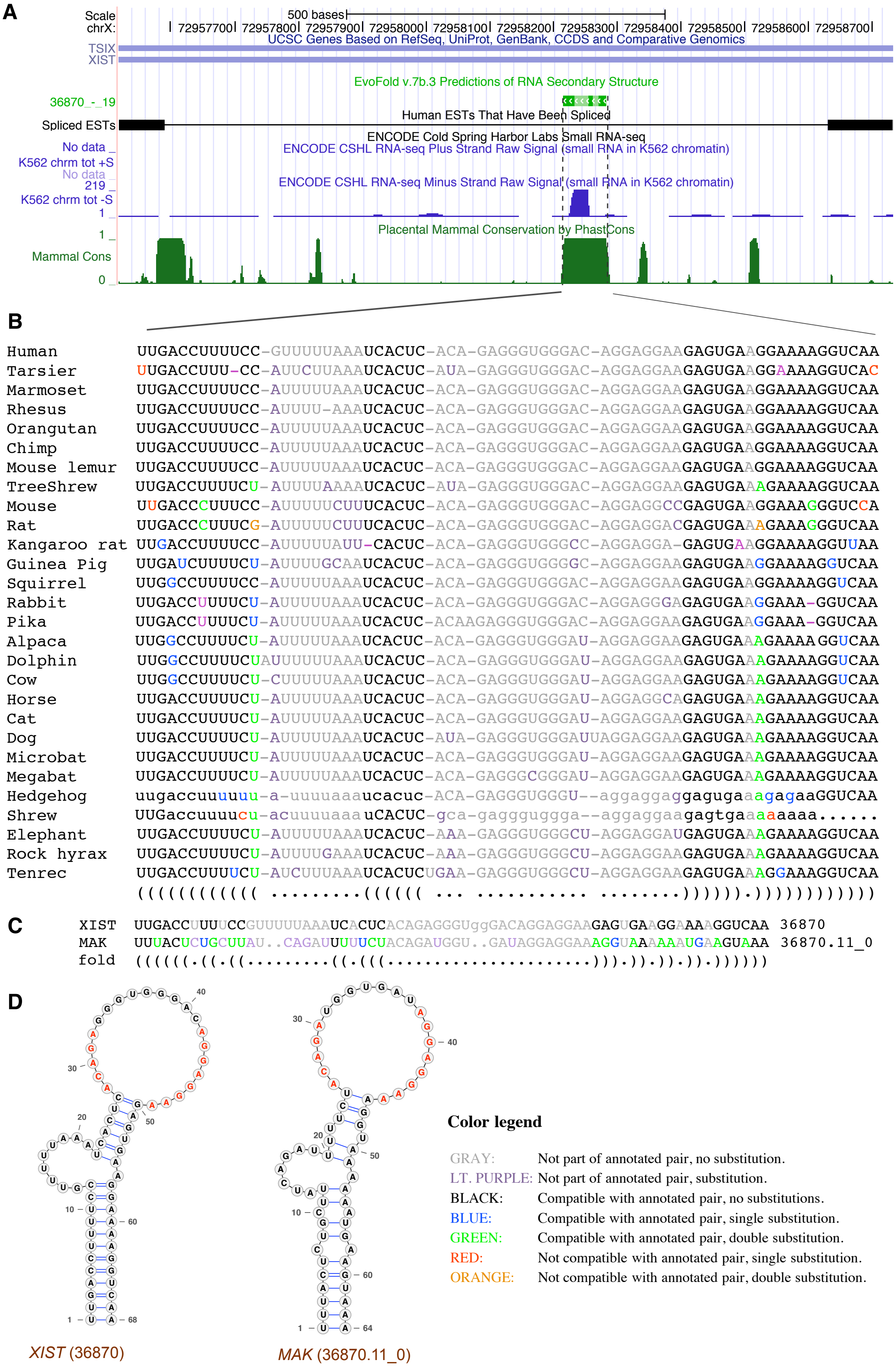

A
Scale
chrX:
100 bases
136795550
136795600
136795650
136795700
136795750
136795800
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
EvoFold v.7b.3 Predictions of RNA Secondary Structure
37234_-_25
ENCODE Cold Spring Harbor Labs Small RNA-seq
ENCODE CSHL RNA-Seq Minus Strand Raw Signal (short in K562 cell)
K562 cell tot -S
ENCODE CSHL RNA-seq Minus Strand Raw Signal (small RNA in K562 nucleoplasm)
K562 nplm tot -S
ENCODE GIS Subcellular RNA Localization by Paired End diTag Sequencing
ENCODE GIS PET Minus Strand Raw Signal Rep 1 (PolyA+ RNA in GM12878 nucleus)
GM12 nucl A+ -S1
ENCODE GIS PET Minus Strand Raw Signal Rep 1 (PolyA+ RNA in K562 nucleus)
K562 nucl A+ -S1
Placental Mammal Conservation by PhastCons
Mammal Cons
B
Scale
chr13:
100 bases
71972550
71972600
71972650
71972700
71972750
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
EvoFold v.7b.3 Predictions of RNA Secondary Structure
9361_+_14
ENCODE Cold Spring Harbor Labs Long RNA-seq
ENCODE CSHL Long RNA-seq Minus Strand Raw Signal Rep 1 (PolyA+ in K562 cytosol)
K562 cyto A+ -S1
ENCODE GIS Subcellular RNA Localization by Paired End diTag Sequencing
ENCODE GIS PET Minus Strand Raw Signal Rep 1 (PolyA+ RNA in GM12878 nucleus)
GM12 nucl A+ -S1
Placental Mammal Conservation by PhastCons
Mammal Cons

**Figure S23: Selected RNA-seq tracks for family GW45.**
The ENCODE CSHL small and large RNA-seq track (Birney et al. 2007) for cell-line K562 for members of family GW45 with A. EvoFold id 37234  and B. EvoFold id 9361 does not show the typical expression profile of a miRNA: 37234 and 9361 show only single cytoplasmic or cellular reads 16 nt and 17 nt, respectively, overlying the stem and loop, which does not match the typical miRNA expression signature of 21-23 nt reads with miR* reads. (B. shows long RNA-seq nominally for  RNA > 200 nt, and so the read may represents a fragment of a larger RNA). There is some evidence of transcription products in the nucleus from the ENCODE paired-end ditag track, with 6x5′ tags across multiple tissues overlapping 37234 and 2x5′ tags overlapping 9361. The expression evidence also suggests that structure 9361 may be located on the minus strand).


A
Scale
chrX:
500 bases
72957700
72957800
72957900
72958000
72958100
72958200
72958300
72958400
72958500
72958600
72958700
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
TSIX
XIST
EvoFold v.7b.3 Predictions of RNA Secondary Structure
36870_-_19
Human ESTs That Have Been Spliced
Spliced ESTs
ENCODE Cold Spring Harbor Labs Small RNA-seq
No data
ENCODE CSHL RNA-seq Plus Strand Raw Signal (small RNA in K562 chromatin)
K562 chrm tot +S
No data
219
ENCODE CSHL RNA-seq Minus Strand Raw Signal (small RNA in K562 chromatin)
K562 chrm tot -S
1
1
Placental Mammal Conservation by PhastCons
Mammal Cons
0
B
C
D
XIST (36870)
MAK (36870.11_0)

```
Human        UUGACCUUUUCC-GUUUUUAAAUCACUC-ACA-GAGGGUGGGAC-AGGAGGAAGAGUGAAGGAAAAGGUCAA
Tarsier      UUGACCUUU-CC-AUUCUUAAAUCACUC-AUA-GAGGGUGGGAC-AGGAGGAAGAGUGAAGGAAAAGGUCAC
Marmoset     UUGACCUUUUCC-AUUUUUAAAUCACUC-ACA-GAGGGUGGGAC-AGGAGGAAGAGUGAAGGAAAAGGUCAA
Rhesus       UUGACCUUUUCC-AUUUU-AAAUCACUC-ACA-GAGGGUGGGAC-AGGAGGAAGAGUGAAGGAAAAGGUCAA
Orangutan    UUGACCUUUUCC-AUUUUUAAAUCACUC-ACA-GAGGGUGGGAC-AGGAGGAAGAGUGAAGGAAAAGGUCAA
Chimp        UUGACCUUUUCC-AUUUUUAAAUCACUC-ACA-GAGGGUGGGAC-AGGAGGAAGAGUGAAGGAAAAGGUCAA
Mouse lemur  UUGACCUUUUCC-AUUUUUAAAUCACUC-ACA-GAGGGUGGGAC-AGGAGGAAGAGUGAAGGAAAAGGUCAA
TreeShrew    UUGACCUUUUCU-AUUUUAAAAUCACUC-AUA-GAGGGUGGGAC-AGGAGGAAGAGUGAAAGAAAAGGUCAA
Mouse        UUGACCCUUUCC-AUUUUUCUUUCACUC-ACA-GAGGGUGGGAC-AGGAGGCCGAGUGAAGGAAAGGGUCCA
Rat          UUGACCCUUUCG-AUUUUUCUUUCACUC-ACA-GAGGGUGGGAC-AGGAGGACGAGUGAAAGAAAGGGUCAA
Kangaroo rat UUGACCUUUUCC-AUUUUUAUU-CACUC-ACA-GAGGGUGGGCC-AGGAGGA-GAGUGAAGGAAAAGGUUAA
Guinea Pig   UUGAUCUUUUCU-AUUUUGCAAUCACUC-ACA-GAGGGUGGGGC-AGGAGGAAGAGUGAAGGAAAAGGUCAA
Squirrel     UUGGCCUUUUCC-AUUUUUAAAUCACUC-ACA-GAGGGUGGGAC-AGGAGGAAGAGUGAAGGAAAAGGUCAA
Rabbit       UUGACCUUUUCU-AUUUUUAAAUCACUC-ACA-GAGGGUGGGAC-AGGAGGGAGAGUGAAGGAAA-GGUCAA
Pika         UUGACCUUUUCU-AUUUUUAAAUCACUC-ACAAGAGGGUGGGAC-AGGAGGAAGAGUGAAGGAAA-GGUCAA
Alpaca       UUGGCCUUUUCU-AUUUUUAAAUCACUC-ACA-GAGGGUGGGAU-AGGAGGAAGAGUGAAAGAAAA GGUCAA
Dolphin      UUGGCCUUUUCUAUUUUUUAAAUCACUC-ACA-GAGGGUGGGAU-AGGAGGAAGAGUGAAAGAAAAGGUCAA
Cow          UUGGCCUUUUCU-CUUUUUAAAUCACUC-ACA-GAGGGUGGGAU-AGGAGGAAGAGUGAAAGAAAAGGUCAA
Horse        UUGACCUUUUCU-AUUUUUAAAUCACUC-ACA-GAGGGUGGGAU-AGGAGGCAGAGUGAAAGAAAAGGUCAA
Cat          UUGACCUUUUCU-AUUUUUAAAUCACUC-ACA-GAGGGUGGGAU-AGGAGGAAGAGUGAAAGAAAAGGUCAA
Dog          UUGACCUUUUCU-AUUUUUAAAUCACUC-AUA-GAGGGUGGGAUUAGGAGGAAGAGUGAAAGAAAAGGUCAA
Microbat     UUGACCUUUUCU-AUUUUUAAAUCACUC-ACA-GAGGGUGGGAU-AGGAGGAAGAGUGAAAGAAAAGGUCAA
Megabat      UUGACCUUUUCU-AUUUUUAAAUCACUC-ACA-GAGGGCGGGAU-AGGAGGAAGAGUGAAAGAAAAGGUCAA
Hedgehog     uugaccuuuuuu-a-uuuuaaaucacuc-ACA-GAGGGUGGGU--aggaggaggagugaaagagaaGGUCAA
Shrew        UUGaccuuuucu-acuuuuaaa uCACUC-gca-gaggguggga--aggaggaagagtgaaaaaaaa......
Elephant     UUGACCUUUUCU-AUUUUUAAAUCACUC-AAA-GAGGGUGGGCU-AGGAGGAUGAGUGAAAGAAAAGGUCAA
Rock hyrax   UUGACCUUUUCU-AUUUUGAAAUCACUC-AAA-GAGGGUGGGCU-AGGAGGAAGAGUGAAAGAAAAGGUCAA
Tenrec       UUGACCUUUUCU-AUCUUUAAAUCACUCUGAA-GAGGGUGGGCU-AGGAGGAAGAGUGAAAGGAAAGGUCAA
             ((((((((((( .........(((((( ... ........... ........))))))).)))))))))))
```

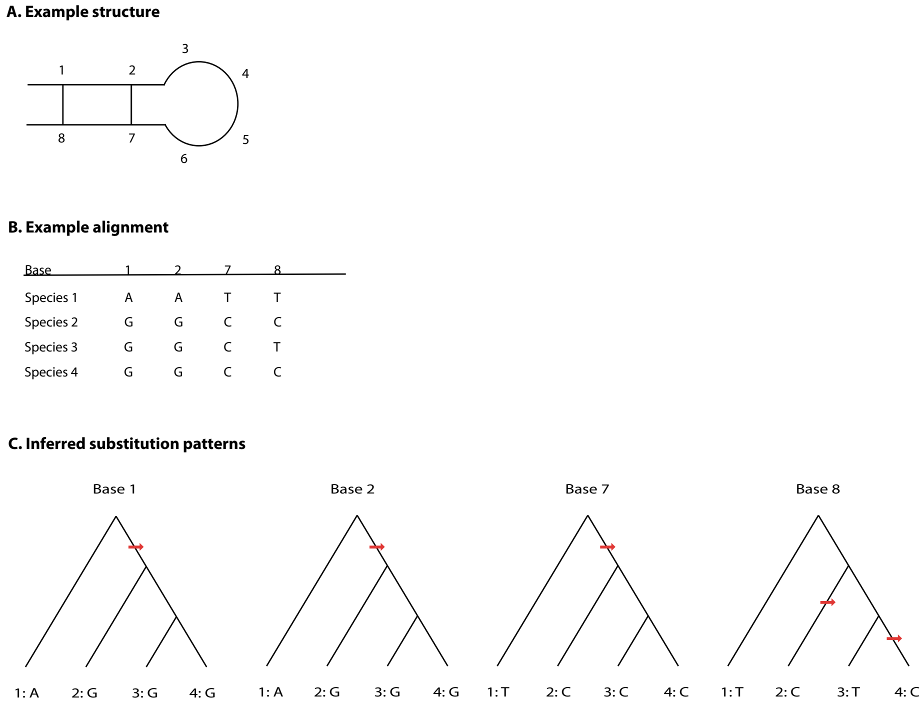
```
XIST  UUGACCUUUUCCGUUUUUAAAUCACUCACAGAGGGUggGACAGGAGGAAGAGUGAAGGAAAAGGUCAA 36870
MAK   UUUACUCUGCUUAU..CAGAUUUUUCUACAGAUGGU..GAUAGGAGGAAAGGUAAAAAAUGAAGUAAA 36870.11_0
fold  ((((((.((.((.........((.(((.......................))).)).)).)).))))))
```

**Color legend**

| | |
|---|---|
| GRAY: | Not part of annotated pair, no substitution. |
| LT. PURPLE: | Not part of annotated pair, substitution. |
| BLACK: | Compatible with annotated pair, no substitutions. |
| BLUE: | Compatible with annotated pair, single substitution. |
| GREEN: | Compatible with annotated pair, double substitution. |
| RED: | Not compatible with annotated pair, single substitution. |
| ORANGE: | Not compatible with annotated pair, double substitution. |

**Figure S24: *XIST* family (EvoFam id: GWP786).**
(A) *XIST* structure genomic location. The structure shows substantial RNA-seq evidence of transcription strongest in the chromatin cellular component (217 uniquely mapped reads). The 5′ ends are predominantly aligned at chrX:72,958,255, possibly a site of transcript cleavage. (B) *XIST* hairpin

multiple species alignment. (C) Alignment of human sequences of family (based on cmalign with *MAK* CM). (D) Structure diagrams. Loop motifs in red.



| Base | 1 | 2 | 7 | 8 |
|---|---|---|---|---|
| Species 1 | A | A | T | T |
| Species 2 | G | G | C | C |
| Species 3 | G | G | C | T |
| Species 4 | G | G | C | C |

**Figure S25: Difference between the two substitution counting approaches.**
In a) a very simple predicted structure is depicted and in b) an example sequence alignment of the stem bases in this structure is shown. Traditionally the number of substitutions is counted by comparing each of the sequences in the alignment to the top (reference) sequence. In this case this gives a substitution count of 11. In the method used for the new p-value the underlying substitution pattern for each base is inferred, see c) and the actual number of substitutions is then counted. In this case this gives a substitution count of 5. One other difference is that traditionally the number of double substitutions is set to the number of times both bases in a base-pair are substituted in the same sequence; here this gives a count of 5 (2 in sequence 2,1 in sequence 3 and 2 in sequence 4). In the new method, we instead count the number of times both bases in a base-pair are substituted in the same sequence *and* on the same branch in the tree. In this case this gives a count of 1 (the base 2-7 GC substitution).

## Supplementary tables:

**Table S1: Top 10 detected GW families.**
Top-10 families, ranked by independent substitution evidence, of the genome-wide (GW) set. See http://moma.ki.au.dk/prj/mammals for raw data files, full set of annotated families, and links to the UCSC Genome Browser.

| **ID** | **Count** | **p-values (EvoP dep./indep.)** | **Mean length (bp)** | **Known functional RNA** | **Description** |
|---|---|---|---|---|---|
| GW1 | 3 | 1.9e-4/1.2E-10 | 27 | mir-103 | *PANK* (intronic) |
| GW2 | 2 | 3.2E-5/2.4E-5 | 13 | snoRNA | *C12orf41* (intronic) |
| GW3 | 2 | 1.0/9.4E-5 | 23 | mir-15 | intronic / intergenic |
| GW4/UTRP1 | 2 | 1.3E-6/1.7E-4 | 16 | | *MAT2A* (3'UTR) |
| GW5 | 2 | 1.0/5.3E-3 | 14 | | *SP4* |
| GW6 | 10 | 5.7E-1/5.4E-3 | 28 | let-7 mir | intronic / intergenic |
| GW7 | 2 | 2.5E-1/2.5E-2 | 15 | | intronic / intergenic |
| GW8 | 2 | 1.0/2.6E-2 | 7 | | intronic / intergenic |
| GW9 | 2 | 4.0E-2/2.7E-2 | 9 | | intronic |
| GW10 | 3 | 1.5E-2/3E-2 | 29 | mir-29/130 | intergenic |

## Table S2: Data set element counts.

| Sets | Filtered | Unfiltered |
|---|---|---|
| GW total | 725 | 3293 |
| GW 5′UTR | 15 | 31 |
| GW 3′UTR | 146 | 446 |
| GW intron | 252 | 1108 |
| GW intergenic | 312 | 1708 |
| UTRP total (paralogs) | 351 (128) | 1589 (714) |
| UTRP 5′UTR | 41 | 181 |
| UTRP 3′UTR | 306 | 1384 |
| GWP total (paralogs) | 2550 (1121) | 14973 (7707) |
| GWP 5′UTR | 77 | 201 |
| GWP 3′UTR | 352 | 1628 |
| GWP intron | 877 | 5644 |
| GWP intergenic | 1160 | 7396 |

## Table S3: Overlap counts of datasets.

| Sets (intersection) | Filtered | Unfiltered |
|---|---|---|
| GW | 725 | 3293 |
| GWP | 2550 | 14973 |
| UTRP | 351 | 1589 |
| GW ∩ GWP | 580 | 2763 |
| GW ∩ UTRP | 100 | 277 |
| GWP ∩ UTRP | 147 | 567 |
| GW ∩ GWP ∩ UTRP | 94 | 242 |

**Table S4: Known cis-regulatory elements detected in families.**
Excludes histone stem loop elements and excludes ncRNAs: *MALAT1*, miRNAs, snoRNAs, tRNAs. Genomic coordinates are of the detected structure.

| EvoFold id | Genomic coords | Strand | Size (bp) | Gene | Description |
|---|---|---|---|---|---|
| 25411 | chr3:197,261,731-197,261,759 | -/- | 10 | TFRC | Iron responsive element (IRE) |
| 25413 | chr3:197,262,296-197,262,326 | -/- | 12 | TFRC | Iron responsive element |
| 25414 | chr3:197,262,245-197,262,277 | +/- | 13 | TFRC | Iron responsive element |
| 25412 | chr3:197,261,843-197,261,871 | -/- | 11 | TFRC | Iron responsive element |
| 25414.5_0 | chr3:197,261,781-197,261,803 | +/- | 8 | TFRC | Iron responsive element |
| 26778 | chr4:158,500,745-158,500,799 | +/+ | 19 | GRIA2 | Editing hairpin |
| 37078 | chrX:122,426,637-122,426,702 | +/+ | 29 | GRIA3 | Editing hairpin |
| 6660 | chr11:105,309,909-105,309,955 | -/+ | 17 | GRIA4 | Editing hairpin |
| 29670 | chr6:31,653,863-31,653,877 | +/+ | 6 | TNF | Constitutive decay element (CDE) |
| 22635.4_0 | chr17:35,427,088-35,427,104 | -/+ | 7 | CSF3 | Stem loop decay element (SLDE) |
| 14986 | chr17:45,633,861-45,633,902 | +/- | 17 | COL1A1 | Collagen 5′ stem-loop |
| 32192 | chr7:93,862,261-93,862,283 | -/+ | 8 | COL1A2 | Collagen 5′ stem-loop |
| 20844 | chr2:189,547,442-189,547,464 | -/+ | 8 | COL3A1 | Collagen 5′ stem-loop |
| 7732 | chr12:46,679,928-46,679,965 | +/- | 13 | COL2A1 | Structured intronic splicing enhancer/silencer |

### Table S5: Recall rates on families of structural RNAs inferred from database annotations.

As explained in the Methods section of the main text, structural RNA annotations were collected from various public databases. Some of these are computationally inferred. In the case of Rfam, human instances were mapped to the human genome based on their primary sequence. The annotations may therefore contain some false positives, especially in the case of large ncRNA families with many pseudo-genes (such as U6). This was accepted as comprehensiveness was a design criteria. The combined set of annotations divided into families based on structural RNA type. miRNAs and tRNAs combined into single families due to uncertainty of sub-grouping. The snoRNAs were split into three families: C/D-box snoRNAs, H/ACA-box snoRNAs, and scaRNAs. All families with more than two instances were included in the table as a family. Due to the problem of false annotations, some families may contain too many members.

The table gives the number of members of each family (Total count); the number overlapping the conserved input regions that are screened (Conserved input); the number overlapping the EvoFold non-protein-coding predictions; and the number found in different EvoFam prediction sets.

| **Name** | **Total count** | **Conserved input** | **EvoFold input** | **EvoFam GW** | **EvoFam GWP** | **EvoFam UTRP** |
|---|---|---|---|---|---|---|
| miRNA | 759 | 431 | 234 | 139 | 155 | 1 |
| tRNA | 473 | 392 | 13 | 2 | 2 | 0 |
| C/D-box snoRNA | 262 | 189 | 9 | 0 | 1 | 0 |
| H/ACA-box snoRNA | 208 | 142 | 25 | 2 | 6 | 0 |
| Histone 3' SL | 67 | 66 | 45 | 45 | 54 | 51 |
| U6 | 45 | 5 | 0 | 0 | 0 | 0 |
| scaRNA | 35 | 23 | 3 | 0 | 2 | 0 |
| Y RNA | 30 | 4 | 0 | 0 | 0 | 0 |
| non-TFRC IRE | 24 | 20 | 0 | 0 | 0 | 0 |
| U1 | 15 | 8 | 0 | 0 | 0 | 0 |
| SECIS | 14 | 13 | 4 | 0 | 0 | 0 |
| U6atac | 6 | 6 | 1 | 0 | 0 | 0 |
| U3 | 5 | 1 | 0 | 0 | 0 | 0 |
| Vault | 4 | 3 | 0 | 0 | 0 | 0 |
| Prion pknot | 3 | 1 | 0 | 0 | 0 | 0 |
| SRP | 3 | 3 | 0 | 0 | 0 | 0 |
| 5S rRNA | 2 | 0 | 0 | 0 | 0 | 0 |
| GAIT | 2 | 1 | 0 | 0 | 0 | 0 |
| IRES Cx43 | 2 | 2 | 0 | 0 | 0 | 0 |
| IRES Hsp70 | 2 | 2 | 0 | 0 | 0 | 0 |
| U4 | 2 | 2 | 0 | 0 | 0 | 0 |
| U4atac | 2 | 1 | 1 | 0 | 0 | 0 |
| U8 | 2 | 0 | 0 | 0 | 0 | 0 |

**Table S6: Recall rates on families of structural RNAs extracted from the literature.** Known families of structural RNAs (all cis-regulatory apart from mascRNA) were extracted from the literature (Koeller et al. 1989; Lomeli et al. 1994; Stefanovic and Brenner 2003; Wilusz et al. 2008; Sunwoo et al. 2009; Wilusz and Spector 2010). The table gives the number of members of each family (Total count); the number overlapping the conserved input regions that are screened (Conserved input); the number overlapping the EvoFold non-protein-coding predictions; and the number found in different EvoFam prediction sets.

| **Name** | **Total count** | **Conserved input** | **EvoFold input** | **EvoFam GW** | **EvoFam GWP** | **EvoFam UTRP** |
|---|---|---|---|---|---|---|
| TFRC IRE | 5 | 5 | 4 | 4 | 5 | 5 |
| COL 5' SL | 3 | 3 | 3 | 3 | 3 | 3 |
| GRIA R/G edit | 3 | 3 | 3 | 3 | 3 | 0 |
| mascRNA type | 2 | 2 | 1 | 0 | 2 | 0 |

## Table S7: Immunity-related families.

All coordinates relative to hg18 assembly; UTRP families shown; Strand column shows predicted/enclosing gene strand; Size column shows number of base-pairs.

(a) UTRP38

| EvoFold id | Genomic coords | Strand | Size (bp) | Gene | Description |
|---|---|---|---|---|---|
| 29670 | chr6:31,653,863-31,653,877 | +/+ | 6 | *TNF* | tumor necrosis factor alpha |
| 37150 | chrX:129,334,838-129,334,852 | +/+ | 6 | *SLC25A14* | solute carrier family 25, member 14 isoform |
| 21066 | chr2:204,446,861-204,446,875 | +/+ | 6 | *CTLA4* | cytotoxic T-lymphocyte-associated protein 4 |
| 10815 | chr14:74,818,492-74,818,506 | -/+ | 6 | *FOS* | v-fos FBJ murine osteosarcoma viral oncogene |
| 29670.8_0 | chr3:161,701,334-161,701,348 | -/- | 6 | *KPNA4* | karyopherin alpha 4 |
| 29762.87_0 | chr3:38,565,442-38,565,456 | +/- | 6 | *SCN5A* | voltage-gated sodium channel type V alpha |
| 13974.10_0 | chr9:126,682,577-126,682,591 | -/- | 6 | *GOLGA1* | golgin 97 |
| 3416.38_0 | chr5:137,303,315-137,303,329 | -/+ | 6 | *PKD2L2* | polycystic kidney disease 2-like 2 |

(b) UTRP36

| EvoFold id | Genomic coords | Strand | Size (bp) | Gene | Description |
|---|---|---|---|---|---|
| 13974 | chr16:83,500,419-83,500,433 | +/+ | 6 | *CRISPLD2* | cysteine-rich secretory protein LCCL domain |
| 24246 | chr3:103,061,006-103,061,020 | -/+ | 6 | *NFKBIZ* | nuclear factor of kappa light polypeptide gene |
| 17704 | chr2:11,883,817-11,883,831 | -/+ | 6 | *LPIN1* | lipin 1 |
| 4494 | chr10:88,673,794-88,673,808 | +/+ | 6 | *BMPR1A* | bone morphogenetic protein receptor, type IA |
| 17320 | chr19:41,071,043-41,071,057 | -/- | 6 | *NFKBID* | nuclear factor of kappa light polypeptide gene |
| 2897 | chr1:203,323,506-203,323,520 | +/- | 6 | *RBBP5* | retinoblastoma binding protein 5 |
| 15563 | chr18:2,907,963-2,907,977 | +/- | 6 | *LPIN2* | lipin 2 |
| 18160 | chr2:47,888,142-47,888,156 | +/- | 6 | *FBXO11* | F-box only protein 11 isoform 1 |

(c) UTRP40

| EvoFold id | Genomic coords | Strand | Size (bp) | Gene | Description |
|---|---|---|---|---|---|
| 10469 | chr14:51,267,058-51,267,074 | +/+ | 7 | *FRMD6* | FERM domain containing 6 |
| 22635 | chr22:27,514,052-27,514,068 | -/+ | 7 | *CCDC117* | coiled-coil domain containing 117 |
| 14833 | chr17:41,464,610-41,464,626 | +/- | 7 | *KIAA1267* | hypothetical protein LOC284058 |
| 8747 | chr12:120,753,855-120,753,871 | +/+ | 7 | *KIAA1076* | homo sapiens mRNA for KIAA1076 protein |
| 30549 | chr6:102,623,141-102,623,157 | +/+ | 7 | *GRIK2* | glutamate receptor, ionotropic, kainate 2 |
| 4990 | chr10:115,796,059-115,796,073 | -/+ | 6 | *ADRB1* | beta-1-adrenergic receptor |
| 29647 | chr6:30,676,288-30,676,304 | +/- | 7 | *PPP1R10* | protein phosphatase 1, regulatory subunit 10 |
| 20134 | chr2:159,884,542-159,884,566 | -/- | 10 | *BAZ2B* | bromodomain adjacent to zinc finger domain, 2B |
| 6457 | chr11:73,389,834-73,389,850 | +/- | 7 | *UCP3* | uncoupling protein 3 isoform UCP3L |
| 2029 | chr1:99,128,588-99,128,604 | +/- | 7 | *PAP2D* | phosphatidic acid phosphatase type 2d isoform 1 |
| 22635.4_0 | chr17:35,427,088-35,427,104 | -/+ | 7 | *CSF3* | colony stimulating factor 3 isoform a precursor |
| 22635.61_0 | chr12:104,086,210-104,086,226 | -/+ | 6 | *KIAA1033* | hypothetical protein LOC23325 |
| 29647.2_0 | chr6:30,639,289-30,639,305 | -/+ | 7 | *PRR3* | proline-rich protein 3 isoform a |

**Table S8: Ion channel gene-related families.**
All coordinates relative to hg18 assembly; Strand column shows predicted/enclosing gene strand; Size column shows number of base-pairs.

(a) GW129

| **EvoFold id** | **Genomic cords** | **Strand** | **Size (bp)** | **Gene** | **Description** |
|---|---|---|---|---|---|
| 32090 | chr7:81,450,660-81,450,696 | +/- | 11 | *CACNA2D1* | calcium channel, voltage-dependent, alpha 2/delta subunit |
| 23518 | chr3:50,385,615-50,385,654 | -/- | 16 | *CACNA2D2* | calcium channel, voltage-dependent, alpha 2/delta subunit 2 |

(b) GW177

| **EvoFold id** | **Genomic cords** | **Strand** | **Size (bp)** | **Gene** | **Description** |
|---|---|---|---|---|---|
| 28742 | chr5:153,155,174-153,155,211 | +/+ | 15 | *GRIA1* | glutamate receptor, ionotropic, AMPA 1 |
| 37081 | chrX:122,441,542-122,441,582 | -/+ | 15 | *GRIA3* | glutamate receptor, ionotropic, AMPA 3 |
| 6668 | chr11:105,347,799-105,347,838 | -/+ | 16 | *GRIA4* | glutamate receptor, ionotropic, AMPA 4 |

(c) UTRP19

| **EvoFold id** | **Genomic cords** | **Strand** | **Size (bp)** | **Gene** | **Description** |
|---|---|---|---|---|---|
| 20387 | chr2:165,955,691-165,955,769 | +/+ | 18 | *SCN2A* | sodium channel, voltage-gated, type II, alpha subunit |
| 20394 | chr2:166,554,937-166,555,003 | -/- | 20 | *SCN1A* | sodium channel, voltage-gated, type I, alpha subunit |
| 20387.2_0 | chr2:165,653,900-165,653,978 | -/- | 18 | *SCN3A* | sodium channel, voltage-gated, type III, alpha subunit |

### Table S9: Members of the ubiquitin-system and immunity enriched family UTRP37/GWP66.

| EvoFold id | Genomic coords | Strand | Size (bp) | Gene | Description |
|---|---|---|---|---|---|
| 33724 | chr8:74,868,530-74,868,547 | -/- | 7 | *UBE2W* | ubiquitin-conjugating enzyme E2W |
| 18028 | chr2:39,066,411-39,066,430 | -/- | 8 | *SOS1* | son of sevenless homolog 1 |
| 6134 | chr11:46,836,503-46,836,520 | -/+ | 7 | *LRP4* | low density lipoprotein receptor-related protein 4 |
| 36885 | chrX:76,649,611-76,649,632 | -/- | 9 | *ATRX* | transcriptional regulator ATRX isoform 1 |
| 8354 | chr12:93,939,541-93,939,558 | -/- | 7 | *NR2C1* | nuclear receptor subfamily 2, group C, member 1 |
| 18028.2_0 | chr16:49,388,001-49,388,020 | +/+ | 8 | *CYLD* | ubiquitin carboxyl-terminal hydrolase CYLD |
| 33724.2_0 | chr5:43,325,347-43,325,364 | -/- | 7 | *HMGCS1* | hydroxymethylglutaryl-CoA synthase 1 |
| 18028.84_0 | chr3:52,411,079-52,411,098 | +/- | 8 | *BAP1* | BRCA1 associated protein-1 |
| 23438 | chrX:10,638,170-10,638,187 | +/- | 7 | *MID1* | midline 1 |
| 8354.4_0 | chr3:47,097,336-47,097,357 | +/- | 9 | *SETD2* | SET domain containing 2 |

### Table S10: Long hairpins clustering with known miRNAs in GW set.

These potentially represent novel miRNAs, other noncoding RNAs, or cis-regulatory structures. Note: 36766 in *CLCN5* has since been added to mirBase v. 15; 16970 and 19405 are members of a size three family (GW169) of structures overlapping likely 5.8s rRNA pseudogenes-- the third member 12937 was classified as miR-1826 in miRBase v. 13 but this annotation has since been withdrawn in miRBase v. 15; 37078 in *GRIA3* is a known A to I editing hairpin (Lomeli et al. 1994).

| **EvoFold id** | **Genomic coordinates** | **Strand** | **Length (bp)** | **Description** |
|---|---|---|---|---|
| 3701 | chr10:11,418,202-11,418,262 | +/+ | 21 | 3'UTR *CUGBP2* |
| 1758 | chr1:88,528,765-88,528,825 | -/na | 23 | |
| 12617 | chr16:5,918,850-5,918,920 | -/na | 29 | |
| 1860 | chr1:90966278-90966325 | +/na | 12 | |
| 35391 | chr9:83,594,513-83,594,617 | -/na | 42 | |
| 1997 | chr1:97,492,862-97,492,947 | -/- | 37 | intron *DPYD* |
| 37116 | chrX:123,865,377-123,865,446 | -/- | 25 | intron *ODZ1* |
| 20584 | chr2:174,944,766-174,944,848 | -/- | 28 | intron *CIR* |
| 30981 | chr6:143,182,515-143,182,592 | +/- | 26 | intron *HIVEP2* |
| 37078 | chrX:122,426,637-122,426,702 | +/+ | 29 | intron *GRIA3* |
| 2023 | chr1:98,342,636-98,342,726 | +/na | 37 | (lincRNA annotation) |
| 36766 | chrX:49,662,030-49,662,086 | +/+ | 19 | intron *CLCN5* |
| 16970 | chr19:23,979,061-23,979,135 | +/na | 17 | |
| 19405 | chr2:132,727,260-132,727,334 | -/na | 18 | |
| 37321 | chrX:151,729,768-151,729,835 | -/na | 21 | |

**Table S11: Distributions of mature read lengths in HeLa and MB-MDA231 breast adenocarcinoma for *CLCN5* putative microRNA.**

The alternative 25 nt isomiR, TAATCCTTGCTACCTGGGTGAGAGT, also shows high reads in several tissues.

HeLa

| **Read length** | 22 | 23 | 24 | 25 |
|---|---|---|---|---|
| **Counts** | 15 | 99 | 7 | 54 |

MB-MDA231 human breast adenocarcinoma

| **Read length** | 22 | 23 | 24 | 25 |
|---|---|---|---|---|
| **Counts** | 29 | 150 | 29 | 153 |

## Table S12: GO enrichment of lincRNAs.

GO analysis was performed on the protein-coding genes of putative families containing lincRNA members in the GWP set. The GO enrichment of the lincRNAs is based on structural similarities and so is a different approach to the analysis of correlated expressions described in (Guttman et al. 2009); the two analyses, however, show some similar results. The highest ranked GO terms in the biological process (BP) ontology was "regulation of regulatory T cell differentiation" (p-value=4.5e-04), c.f. (Guttman et al. 2009) where it was noted that one cluster of lincRNAs had immunity/inflammation enrichment. The histone methyltransferase enrichment terms in GO metabolic function (MF) may similarly be compared with the chromatin-modification role of lincRNAs noted in (Khalil et al. 2009). Several enrichments appear to correspond to a cellular adhesion function including "proteoglycan metabolic process" (p-value=1.0e-03) in BP and "heparan sulfate sulfotransferase activity" (2.8e-03) in MF (proteoglycans comprise part of the extracellular matrix, with the common cell surface component heparin sulfate, and are involved in cell adhesion and related processes); and GO cellular content (CC) ontology similarly showed enrichment for cell adhesion-related terms including "apical junction complex" (p-value=2.1e-03). The full list of GO terms with p-value < 1E-2 follows:

| **GO biological process (BP) term** | **p-value** |
|---|---|
| regulation of regulatory T cell differentiation | 4.5e-04 |
| proteoglycan metabolic process | 1.0e-03 |
| nucleobase, nucleoside and nucleotide metabolic process | 3.1e-03 |
| transcription from RNA polymerase III promoter | 3.5e-03 |
| heparan sulfate proteoglycan metabolic process | 3.9e-03 |
| regulation of TOR signaling pathway | 4.3e-03 |
| establishment and/or maintenance of epithelial cell apical/basal polarity | 4.3e-03 |
| nucleotide biosynthetic process | 4.8e-03 |
| establishment and/or maintenance of apical/basal cell polarity | 5.1e-03 |
| sensory perception of smell | 5.6e-03 |
| defense response | 5.8e-03 |
| sensory perception of chemical stimulus | 6.0e-03 |
| cellular response to reactive oxygen species | 6.1e-03 |
| nucleoside phosphate metabolic process | 6.3e-03 |
| nucleotide metabolic process | 6.3e-03 |
| regulation of immunoglobulin secretion | 6.3e-03 |
| TOR signaling pathway | 6.3e-03 |
| cellular response to oxidative stress | 7.1e-03 |
| cell aging | 7.5e-03 |
| regulation of cyclic nucleotide biosynthetic process | 7.5e-03 |
| regulation of nucleotide biosynthetic process | 8.2e-03 |
| cyclic nucleotide metabolic process | 8.7e-03 |
| Vitellogenesis | 8.7e-03 |
| myoblast maturation | 9.9e-03 |

| **GO molecular function (MF) term** | **p-value** |
|---|---|
| phosphatidate phosphatase activity | 2.3e-03 |
| [heparan sulfate]-glucosamine 3-sulfotransferase 1 activity | 2.6e-03 |
| histone methyltransferase activity | 3.1e-03 |
| N-methyltransferase activity | 3.5e-03 |
| protein methyltransferase activity | 4.0e-03 |
| *S*-adenosylmethionine-dependent methyltransferase activity | 6.2e-03 |
| lysine N-methyltransferase activity | 6.4e-03 |
| protein-lysine N-methyltransferase activity | 6.4e-03 |
| histone-lysine N-methyltransferase activity | 6.4e-03 |
| histone lysine N-methyltransferase activity (H4-K20 specific) | 8.9e-03 |

| **GO cellular components (CC) term** | **p-value** |
|---|---|
| proton-transporting V-type ATPase, V0 domain | 4.8e-04 |
| apical junction complex | 2.1e-03 |
| apicolateral plasma membrane | 2.7e-03 |
| tight junction | 8.9e-03 |

## Supplementary results:

**Other uncharacterized families of hairpins with strong independent evidence:**

***XIST* family GWP786:**

The well-known lincRNA *XIST,* which is involved in X chromosome inactivation is a member of the size-two family (EvoFam id: GWP786). It shows an 18 bp hairpin at the 3′ end of *XIST* with strong double substitution evidence (p-value=0.01; EvoP (dependent set)) and 12 compatible single substitutions (genomic coordinates chrX:72,958,217-72,958,284) (see supplementary figure S24). This structure is distinct from the known 5′ structures discussed in (Maenner et al. 2010). *roX2* RNA, which has a related function in Drosophila, is also known to have structured regions involved in its function (Park et al. 2007). X chomosome inactivation requires homologous chromosome pairing at a region 3′ to XIST in mouse (Xu et al. 2006; Xu et al. 2007). Also, *XIST* is known to be post-transcriptionally regulated, possibly through splicing regulation (Ciaudo et al. 2006). There is spliced EST and RNA-seq evidence that this structure is within an unannotated intron. There are 217 short reads uniquely mapped to this structure within the chromatin fraction of cell line K562, with very few reads in the neighbouring region (ENCODE CSHL short RNA-seq) (see figure S24).

The second member of the family was detected as a paralogous match and is a 15-bp structure in the intron of *MAK* (male germ cell-associated kinase) (chr6:10,876,235-10,876,298), which is a serine-threonine kinase expressed almost exclusively in the testis, and localizes to the synaptonemal complex involved in homologous chromosome pairing during meiosis. This structure however shows only weak conserved structural evidence with few compensatory double substitutions (p-value=0.01; EvoP) and also shows some sequence homology in the loop (the motif AGGAGGAA is highly conserved in both members). The binding of *XIST* is known to be regulated by the serine-threonine kinase Aurora B (AURKB) (Hall et al. 2009), which is also involved in chromosome pairing (Lampson et al. 2004). There is some transcript evidence for a totally intronic noncoding RNA overlying the *MAK* structure (Nakaya et al. 2007) with ESTs N72578 and BX119802 from 20-week male fetal tissue. Interestingly, we note that both *XIST* and *MAK* overlap antisense transcripts, *TSIX* and *TMEM14B* respectively.

**Ubiquitin system-enriched family UTRP37/GWP66:**

An additional family showing significant enrichment for macrophage-related genes (4 genes, p-value=1.6E-4; Fisher exact test) is a size 8 family of hairpins (UTRP37;GWP66) with mean length 8 bp, and strong compatible single substitution evidence. This family differs substantially from the other immunity-related families discussed previously: the hairpins show a 4-nt loop and it shows an enrichment for genes involved in protein post-translational modifications, such as those involving the ubiquitin-proteosome system (GO enrichment for proteolysis (p-value=9E-4; Fisher exact test) and ubiquitin thiolesterase activity (p-value=1.1e-03; Fisher exact test)).

Structures are located in the 3′UTRs of four genes directly involved in the ubiquitin pathway: *BAP1*, *CYLD*, *UBE2W*, and *MID1* ((Nishikawa et al. 2009) (Hussain et al. 2009; Yang et al. 2009) (Christensen et al. 2007); (Trockenbacher et al. 2001) (See supplementary table S4).
Across the datasets, 3′UTR members are found in: *BAP1*, *CYLD*, *UBE2W*, *NR2C1*, *SOS1, LRP4/CR612190,* and intronic members are found in: *MID1* and *SETD2.*

***BCL11A/B* family UTRP2:**

Another 3′UTR cluster (UTRP2) consists of *BCL11A* and *B,* which are two transcription factors that have been shown to be required in lymphoid development and *BCL11A* is associated with lymphoid malignancies (Satterwhite et al. 2001). These paralogous genes each have a 9-10 bp long hairpin located at the 3′ end of the 3′UTR, which have strong independent double substitution evidence (p-value=3e-4; EvoP (independent set)).

**Ion-channel family GW59:**

Another family of ion channel-related genes is a family consisting of *CACNB4*, *KCNMA1*, and *ANK3*. *CACNB4* and *KCNMA1* are part of voltage-dependent calcium channel complexes and voltage and calcium-sensitive potassium channel proteins, respectively. *ANK3* is a structural transmembrane protein that is known to regulate neuronal excitability both by clustering voltage-gated sodium channels and modifying their channel gating (Shirahata et al. 2006). This family consists of 3′UTR and intronic short (6-8 bp) hairpins and so is less likely to represent editing targets. Subcellular localization of mRNAs is an important mechanism in neuronal cells (Holt and Bullock 2009). The family shows GO cellular component enrichment for axon (p-value=1.7e-3; Fisher exact test) and synapse (p-value=7.7e-3; Fisher exact test), suggesting possible involvement in localization.

**Alternative splicing putative family GWP413:**

A known example of an intronic structural motif directing alternative splicing (Faustino and Cooper 2003) is in collagen gene *COL2A1* (McAlinden et al. 2005). This corresponds to a member (EvoFold id 7732) of GWP413 (with the other member however being labeled as an intergenic match).

## Supplementary methods:

### Genome-wide EvoFold screen for structural RNAs

A genome-wide input set of structural RNA predictions was made by screening the conserved segments of the 31-way vertebrate alignments using EvoFold (v.2.0) (Pedersen et al. 2006). The 31-way input alignment set was predominantly composed of mammals (see figure S1 and S2) to ensure the strongest signal to detect mammalian structures, as part of the 29 Mammals Sequencing and Analysis Consortium. A further 10 species consisting predominantly of non-mammalian vertebrates was used as an independent test set.

The conserved segments were defined from PhastCons conserved elements extended by 25 bases on both sides and combined (Siepel et al. 2005). These segments were divided into 150 long windows, each overlapping by 50 bases. Before applying EvoFold to both strands of the windows, sequences likely to be mis-aligned or contain sequencing errors were removed using the alignment filtering method described below.

Predictions were first dissected into individual nested structures (folds) as described in (Pedersen et al. 2006). The predictions were quality filtered by removing low confidence base-pairs (posterior probability < 50%), short predictions (< 6 base-pairs), and structures with excessive amount of bulges (more than 40% of bases in stems found in bulges).

Predictions were also discarded if based on shallow or low quality alignments, as follows: (1) 31-way alignment segments corresponding to each structure prediction were extracted. (2) Individual sequences from these alignments were discarded if they contained: more than 15 insertions relative to human, more than 50% missing data, more than 5% lower (repeat) masked sequence, more than 25% of the annotated base-pairs could not form, or more than 7.5% of the base-pairs involved gap characters. (3) If less than eight sequences remained in the alignment or if the human sequence was missing after these filtering steps, the prediction was discarded. These filters are empirically determined by properties of true structural RNAs and have been used successfully previously (Pedersen et al. 2006; Stark et al. 2007).

We also removed predictions overlapping either repeats (as defined by repeatmasker) or pseudogenes (as defined by UCSC Retrogenes or Yale Pseudogene Database (Karro et al. 2007)) by more than 20%. We also flagged predictions that showed strong homology with the mitochondrial chromosome, as these are likely to represent either ribosomal RNA or tRNA pseudogenes, and these were excluded from the filtered sets.

Finally, the remaining predictions were reduced to a non-overlapping set by successively selecting the highest scoring prediction, according to the length-normalized EvoFold log-odds score, when overlap occurred.

### Similarity measure between models of structural RNAs

The profile SCFG models define a probability distribution over the set of possible sequences, with sequences likely to be generated by the model having high probability: they are generative models, modeling insertions and deletions probabilistically, as well as base and base-pair frequencies. Let $M$ be a pSCFG model that generates RNA sequences over the alphabet $\Sigma = \{C,G,A,U\}$ with probability distribution $P_M$, then $P_M(s)$, the probability of sequence $s \in \Sigma^*$ under model $M$, can be computed efficiently by the inside algorithm, using a dynamic programming implementation (Eddy and Durbin 1994; Durbin 1998), and the pSCFG model can be used generatively to sample sequences from this probability distribution.

The similarity measure used is derived from the Kullback-Leibler divergence $D_{KL}$ between probability distributions, which defines a fundamental measure of similarity between the probability distributions $P_{M_1}$ and $P_{M_2}$ of sequences generated by models $M_1$ and $M_2$ respectively,

$$D_{KL}\left(M_1 \,\|\, M_2\right) = \sum_i P_{M_1}(i) \log \frac{P_{M_1}(i)}{P_{M_2}(i)} \tag{1}$$

summed over the set of all $i \in \Sigma^*$ sequences generated by model $M_1$, with $P_{M_{\{1,2\}}}(i)$ the probability of the respective models generating that sequence.

This divergence can be estimated by a Monte Carlo approach (Juang and Rabiner 1985), comparing sequences emitted from the generative model $M_1$ with the probability that they were produced by model $M_2$ using eq. 1.

$$\tilde{D}_{KL}\left(M_1 \,\|\, M_2\right) = 1/n \sum_{i=1}^{n} 1/l(s_{1,i}) \cdot \left(\log\left(P_{M_1}(s_{1,i})\right) - \log\left(P_{M_2}(s_{1,i})\right)\right) \tag{2}$$

for $n$ Monte Carlo samples, where $s_{1,i}$ is the $i$th sequence generated by model 1. In the limit of many samples $n$ this will approach the true Kullback-Leibler divergence.

(Juang and Rabiner 1985) examined hidden Markov models of simple sequences and to make the divergences comparable normalized by the length $l$ of each $s_{1,i}$.

However, there are two limitations of this approach for this study: (1) For a large genome-wide study, such sampling many times over the full probability space is computationally intractable. and (2) the simple length normalization is insufficient for more complex SCFG models of varying structural complexity but perhaps similar length.

To limit computational costs in this genome-wide study, we used an approximation based on using only a single sample from the probability distribution of sequences from model $M_1$. This sample consisted of the human sequence used to train model $M_1$ (which is at the approximate mode of the distribution and the sequence of major interest) to give

$$\tilde{D}_{KL,human}\left(M_1 \,\|\, M_2\right) = 1/l(s_{1,human}) \cdot \left(\log\left(P_{M_1}(s_{1,human})\right) - \log\left(P_{M_2}(s_{1,human})\right)\right) \tag{3}$$

, where $s_{1,human}$ is the human sequence used to train model 1, and normalized by the length $l$ of $s_{1,human}$.

The score $S$ calculated for each alignment by the Infernal tools is returned as the log odds relative to a null model of a random unstructured sequence with

$$S(s,M) = \log_2 \frac{P_M(s)}{P_{null}(s)}. \tag{4}$$

giving

$$\tilde{D}_{KL,human}\left(M_1 \,\|\, M_2\right) = 1/l(s_{1,human}) \cdot \left(S(s_{1,human}, M_1) - S(s_{1,human}, M_2)\right). \tag{5}$$

However, model complexity varies greatly between RNA structures and in genome-wide studies the rate of false positives would be larger for smaller and less complex models (i.e. simple unstructured sequences versus complex looped structures). We desire a similarity measure that is normalized for the expected false positive rate. The length normalization incorporated in $\tilde{D}_{KL,human}$ above is not sufficient to fully correct for these differences in model complexity. Therefore we use a (dis)similarity measure based on E-value i.e. the expected number of false positives with the same or greater score when searching with model $M$ against a database of sequences of total length approximately equal to the combined sequence length searched in performing the all-against-all cluster analysis: $E_M(S)$ denotes the estimation of the E-value corresponding to a given score $S$ when searching using a model $M$. This transformation of scores $S$ to E-value is computed empirically (using the Infernal toolkit), separately for each model $M$, by generating a histogram of scores by searching randomly generated sequences, and then fitting a smooth exponential curve to the tails. $E$ is highly correlated with $S$ but is normalized for differing lengths and complexities of models. This is due to its being based on statistical significance, with the database size $E$ it is computed relative to being identical for all models. Using $E$ rather than length to normalize in the above, the (dis)similarity score used is:

$$\tilde{D}_{E,human}\left(M_1 \,||\, M_2\right) = E_{M_2}\left(S(seq_{1,human}, M_2)\right) - E_{M_1}\left(S(seq_{1,human}, M_1)\right). \qquad (6)$$

The Kullback-Leibler divergence is an information theoretic measure giving the expected number of extra bits required to code samples from $M_1$ when using a code based on $M_2$, rather than using the ideal code based on $M_1$. In comparison with this definition of KL divergence, this new divergence measures the expected number of extra false positives over a database of sequences generated from $M_1$ when searching using model $M_2$, rather than using the ideal model $M_1$, and can be considered a normalized derivative of the KL divergence. Identical models will have a value of 0, and more dissimilar models will have larger values, as for the KL divergence (note that for the single sample approximation described in eq. 6, $\tilde{D}_{E,human}\left(M_1 \,||\, M_2\right)$ will similarly tend to be positive, with smaller values for more similar models; values < 0 are set to 0).

These measures are not symmetric: $\tilde{D}_{E,human}\left(M_1 \,||\, M_2\right)$ is, in general, $\neq \tilde{D}_{E,human}\left(M_2 \,||\, M_1\right)$.

By applying this measure reciprocally, we get an estimate of how likely the two models are to generate the same sequences i.e. how similar the models are. To generate a final symmetric measure we use:

$$D\left(M_1 \,||\, M_2\right) = \max\left(\tilde{D}_{E,human}\left(M_1 \,||\, M_2\right), \tilde{D}_{E,human}\left(M_2 \,||\, M_1\right)\right). \qquad (7)$$

This gives a conservative lower bound on the divergences between the two models. Note that when searching with a model $M$ in this study we use a global alignment of $M$ to the sequence (using cmsearch -g from the Infernal package) to tradeoff sensitivity versus specificity. The most significant hit from searching both strands is used in the above, as the predicted strand from EvoFold is known to have a high error rate. Note that while the search is global relative to the model, it is effectively local relative to the sequence itself. This could lead to e.g. a small hairpin model matching with good score a small part of a larger structure in the sequence, but the maximum symmetrization of $D\left(M_1 \,||\, M_2\right)$ will correctly penalize this case.

**Graph-based family definition**

A similarity graph $G(V,E)$ was defined with vertex set $V$ corresponding to pSCFG models of RNA structures, with edges connecting elements with dissimilarity $D$ below a threshold $T$ . Parameter $T$ was specified to vary the sensitivity/specificity tradeoff for inclusion of edges, to limit the false discovery rate (set to 0.25, 0.25, 1.0 for GW, GWP and UTRP sets respectively). The similarity graph is sparse and was implemented in a space-efficient edge list data structure to allow whole-genome analyses to be efficiently performed.

Families were defined as highly connected subgraphs $S \subset G$, where a highly connected subgraph (HCS) is defined as a subgraph of $n$ vertices with edge connectivity $k(S) > n/2$, where edge connectivity $k(S)$ is defined as the minimum number of edges whose removal disconnects $S$. These families are computed using the iterated HCS algorithm of (Hartuv and Shamir 2000; Carey et al. 2010). Ideally, functional families would appear as cliques i.e. each member would be connected to each other. In practice, the definition of a highly connected subgraph as used here allows partial cliques to also be detected which allows families to be robustly identified in the presence of noise and conservative error rate control. Each such family can be shown to be at least half as dense as a clique (Hartuv and Shamir 2000) ensuring that the families have adequate evidential support. For example, in figure 1C, the yellow subgraph, whilst not being a clique, fulfills the weaker requirements of a HCS. Note that isolated 2-element clusters are explicitly allowed (a 2-element cluster is an HCS by definition), and as shown in figure 1D such small families are common.

**Substitution evidence evaluation (EvoP test)**

The prediction methods we use are based on complex probabilistic models and their scores are subject to various biases. We have therefore developed an easily interpretable p-value for evaluating and ranking the predicted structures.

*Basic definition of the new p-value*

For a given predicted structure *s* and a multi-species sequence alignment of the stem bases of *s*, the new p-value is simply a measure of how probable it would be to see at least as many double substitutions in the aligned sequences if they did not encode a structural RNA (assuming the total number of substitutions in the predicted stem bases is fixed). A very similar approach was taken in a study by Pollard et al. (Pollard et al. 2006), which we were involved in. However, in that study the substitution counts for their p-values were based on comparing each of the aligned sequences to a single reference sequence and a double substitution was simply defined as two substitutions that occur in the same base-pair in the same sequence. To take into account the tree-like nature of evolution we instead perform the substitution counts on the underlying phylogenetic tree of the species in the alignment. And to reflect the fact that the two substitutions are only likely to be correlated if they happen in close proximity time-wise, we define a double substitution as two substitutions that occur not only in the same base-pair in the same sequence, but also on the same branch in the tree. The difference between the two counting approaches is illustrated in figure S25.

When calculating the p-values we assume that if the aligned sequences do not encode a structural RNA, then each substitution has happened with equal probability along the sequence, with probabilities proportional to the branch lengths along the phylogenetic tree and with equal probabilities to all different types of substitutions.

In the cases where the structure prediction is only based on a subset, $S$, of the species in the alignment we use almost the same approach to assess to what extent the additional species support this prediction. The only difference in the approach is that in this case only the branches that do not connect the species in $S$ are included in the analyses.

*Implementation*

Counting substitutions on a tree**:** To count substitutions on a tree we first use the algorithm of Pupko et al. (2000) to infer the most likely ancestral nucleotide sequence in each node in the tree. We then simply count the number of substitutions on each branch by counting the number of differences between the nucleotide sequences in the two nodes that the branch connects.

We use the implementation of Pupko's algorithm that is available in the software package PAML (Yang 2007) and Jukes Cantor as the underlying mutation model.

Estimating the p-value: Unfortunately even for the simple null model we use, the exact p-values are very time-consuming to calculate. We therefore estimate them using a Monte Carlo approach instead. Per definition the p-value of a predicted structure *s* given a multi-species alignment *a* of the *b* stem bases of *s* and a phylogenetic tree *t* is equal to:

$$p = P(D \geq d \mid N = n, B = b, T = t) \tag{8}$$

where *d* is the number of double substitutions in *a* and *n* is the total number of substitutions in *a*. If we let *X* be the set of all possible substitution patterns given *n*, *b* and *t* and let *f* be an indicator function that returns 1 if a pattern *x* has *d* or more double substitutions and 0 otherwise this can be rewritten as

$$p = P(D \geq d \mid N = n, B = b, T = t) = \sum_{x \in X} f(x) p_{null}(x) \tag{9}$$

Hence if we let **X** be a stochastic variable that can take values in *X* and that is distributed according to the null model then the p-value can be formulated as the expectation:

$$p = P(D \geq d \mid N = n, B = b, T = t) = \sum_{x \in X} f(x) p_{null}(x) = E_{null}(f(\mathbf{X})) \tag{10}$$

Thus we can estimate the p-value by a standard Monte Carlo approach for estimating expectations: sample *m* substitution patterns, $x_1, x_2, x_3, \ldots, x_m$ from the null model and then use the estimator:

$$\tilde{p} = \frac{1}{m} \sum_{i=1}^{m} f(x_i) \tag{11}$$

This estimator is unbiased and its variance is Var(f(**X**))/*m*. Hence the larger *m* is the more accurate the estimate will be. We use an *m* value of 10,000,000. However, we do the sampling in batches of 10,000 and stop sampling if we reach 50 samples that are more extreme than the one we are testing. In this way we avoid wasting time on estimating large p-values, but still achieve high accuracy on low p-values; the p-value we are interested in.

### Alignment filtering

The 2x mammalian genome assemblies contain regions covered by only single reads, which therefore have elevated error rates. Since EvoFold identifies conserved RNA structures based on their characteristic substitution pattern, it is sensitive to such errors. Therefore, we filter sequences that are likely to be misaligned away from the alignments before making any predictions. For each alignment, we first identify every outer branch in the underlying phylogenetic tree on which there are significantly more substitutions than expected given its length and the substitution rate in the rest of the tree. More specifically, every outer branch with at least 2 mutations that has a substitution rate which is more than five times larger than the substitution rate in the rest of the tree is identified. If such a branch exists, we mark the entire sequence from the species to which this branch leads as unobserved in the alignment. Both the number of substitutions on the outer branches and the substitution rate in the remaining tree are estimated based on the most probable ancestral nucleotide sequences, which are inferred using the software package PAML [9], see above for details.

**Thermodynamic analysis of structure families with RNAz**

We used RNAz 2.0 (Washietl et al. 2005; Gruber et al. 2010) to analyze the initial EvoFold predictions as well as the structures in the clustered families before and after filtering. RNAz was run on the same alignments as EvoFold. However, since the classification algorithm of RNAz is not trained for short structures, we added flanking regions to all EvoFold predictions shorter than 120 nt to obtain a minimum length of 120.

As positive control we used a set of 356 known structural RNAs. As negative controls we (i) chose random locations within the PhastCons conserved regions that were used as input for the EvoFold analysis, (ii) we shuffled the alignment (Washietl and Hofacker 2004), (iii) we simulated random alignments preserving the dinucleotide content in the alignment (Gesell and Washietl 2008).

We focus on two metrics calculated by RNAz: the z-score and the classification score. The z-score is a normalized value measuring the thermodynamic stability of an RNA structure. It is the number of standard deviations a given RNA structure is more/less stable than structures for random sequences of the same length and dinucleotide content. RNAz calculates the average z-score of all sequences in an alignment. By convention, negative z-scores denote more stable structures. Figure S6 A shows the cumulative distribution of the z-score in all tested sets. The negative control sets have – as expected – an average z-score of ~0, while all other sets are clearly shifted towards negative (i.e. more stable structures) which is characteristic for biologically relevant RNA structures (see for example (Washietl and Hofacker 2004)). Distributions are shifted strongest for the high-confidence set (filtered) families which show similar distributions as the positive control set. This confirms that the clustering procedure and the filters enrich for high confidence structures. It is important to note that the location-, GO- and double substitution- filters are all independent of the stability z-score. Only the length filter is related to the structure. Although the z-score is length normalized, we included a set without the length filter to exclude any possible confounding issues in this statistics.

RNAz combines the stability z-score with an evolutionary score measuring the structural conservation. This combination score is calculated via a support vector machine classifier. Alignments that are classified as "RNA" have scores >0 while all other alignments have scores <0. Also this combined score shows similar enrichment of predicted structure families over random controls and raw EvoFold predictions (figure S6 B). Figure S6 C shows the fraction of structures that are classified as RNA, ranging from 1-2% in the negative controls to 52.2 in the positive set of true structures. The raw EvoFold predictions are in the lower end of this spectrum at 9.7%, while the filtered families are close to the true structures in both the genome-wide and UTR sets with 30%-40% of structures predicted as RNA.

**Computational validation**

*Known families detected.*

A comprehensive set of known structural RNA annotations were extracted from various databases, including Rfam (see Methods in main text). A known structure was counted as detected when a structure prediction overlapped by any number of bases. There are several reasons for using this criterion. First, EvoFold is designed to identify the consensus part of a structure and will generally not detect portions that deviate between species. Neither will purely sequenced-based flanks be detected. Therefore, in some cases only a conserved core is detected and some large structures are broken up into several disjoint EvoFold predictions. Second, the input EvoFold predictions span a very small fraction of the genome (0.034%), which makes chance overlaps rare.

There is no standard set of human structural RNA families that we could benchmark our family predictions against. We therefore defined approximate families based on the types of structural RNAs found among the collected comprehensive set of structural RNA annotations (See Table S10). However, this approach only allows high-level family definition and is of limited accuracy in many cases. For instance, miRNAs are defined as a single family, as the precise evolutionary sub-division among these is not well defined in the available annotation. The family division of miRNAs is based on primary sequence comparisons (often seed comparisons) and thus depends on the chosen thresholds. Similarly, snoRNAs were divided into three main families: C/D-box snoRNAs, H/ACA-box snoRNAs, and scaRNAs. Furthermore, since our comprehensive set includes both human Rfam family members mapped back to the human genome as well as existing computational annotations, we expect a certain level of falsely included pseudo-genes, especially for some types of structural RNAs (e.g., U6 RNA).

As part of the detailed analysis of the predicted families, we extracted some known families from the literature. These well-defined families are explicitly discussed in the main text and we report the recall rate on each (See Table S11).

Because of these limitations of the family annotation, we could not in general reliably estimate the recall rate (sensitivity) for most known families recovered by our screen. We additionally report an estimate of overall precision (positive predictive value): for each predicted family with any known members, as a lower bound estimate we consider all unknown members as false positives. The estimate of precision is thus given by the percentage of known (true positive) elements in this set.

*FDR estimates.*

FDR estimates for the family identification were computationally estimated by comparison with null sets generated by permutation. These null sets maintained the same size and structural distribution as the unshuffled input sets.

To estimate the FDR at the level of individual pairs of similar structures, the original multi-species alignments were randomly shuffled and profile SCFG models retrained from these. The single-stranded and double-stranded regions of the RNA model were shuffled separately, with the double-stranded shuffling keeping base-pairs intact such that the RNA structure remained unchanged. Thus this null model has random sequences but the same length and structural complexity distribution as the original data. The similarity graph construction stage of EvoFam was run on this randomized data, and the ratio of the number of edges in this similarity graph compared with that of the original unshuffled data was used to estimate the pair-wise FDR. Note that repetitive sequence composition in the human genomic background was not explicitly modeled by this null distribution and so false positives due to detected homology between such regions are not included in the FDR estimate, and so the FDR could be higher in practice due to this effect.

Next, we estimated how likely the clustering stage was to produce spurious groupings of these edges. To estimate this family-wise FDR, a randomized null distribution was generated by shuffling the original similarity graph to create a random similarity graph over the original number of vertices. The rate of false positive families ≥ size 3 generated by the clustering method depends upon the density of the background edges not part of such larger families, so a random null model approximating this was created by randomly shuffling the background isolated edges (size two families) of the original unshuffled similarity graph (GW set) amongst the original vertex set. The same edge thresholds used in the unshuffled pipeline were used. Shuffling was performed ensuring that no multiple edges between vertices were created. The shuffled similarity graphs were then clustered. This was repeated 20 times and mean cluster counts computed. Note that due to the computational cost, this analysis was run on the UTRP set only. The ratio of the mean counts of the clusters of each size in the shuffled sets compared with the counts in the unshuffled set were used to estimate the FDR for each family size.

*Overlap with DNaseI hypersensitive regions.*
DNaseI hypersensitive regions represent areas of open chromatin and correlate with functional regions. The predictions in the GW set were overlapped with the University of Washington ENCODE Digital DNaseI Hypersensitivity Clusters track (Birney et al. 2007). To compute a p-value, a permutation test was performed, whereby the overlap was compared with a randomly shuffled set of predictions with the same size distribution as the original set, but drawn from intergenic regions overlapped with the conserved regions of the genome, as defined for the input to EvoFam. 1000 random shuffles were used. For comparison, the results were also compared with those using shuffled predictions drawn from the entire intergenic regions (conserved and non-conserved) of the genome.

*Structural and sequence similarity within paralogous families.*
The overall structural similarity for each family in the GW unfiltered set was calculated as the mean percentage of matching base-pairs (number of mismatched base-pairs/total number of base-pairs), comparing each member against the member structure with the strongest evolutionary support, i.e. smallest EvoP (ordered by independent then dependent set p-values) (see supplementary data S14). Similarly, pairwise sequence similarity was computed for each family. Percent pairwise alignment identity was calculated for each family as the mean, over all pairs of human sequences, of (*identities* / min(*length1*, *length2*)) with *identities* = the number of exact identities; *length1* and *length2* = unaligned length of the two sequences of each pair.

## Expression analyses

*Expression enrichment of intergenic and intronic structures.*
Using the Illumina Body Map 2 ribo-depleted, non-polyA selected, total RNA dataset, pooled for 16 tissues (Illumina Inc. 2010), we compared the mean log expression evidence (mean of $\log_2$(reads/base+1 pseudocount)) overlying the novel intergenic and intronic structures compared with a shuffled set of random structure positions chosen from the conserved intergenic and conserved intronic regions of the genome, respectively. Structures showing mitochondrial chromosome homology were removed as they have increased probability of representing either ribosomal RNA or tRNA pseudogenes. Similarly, mitochondrial chromosome homologous regions, repeat elements and known pseudogenes were removed from the genomic backgrounds. 1000 shuffles were used to estimate p-values by permutation test.

*Expression correlation within families.*
To estimate whether the detected families show an increased correlation of tissue-specific regulation compared with a random background, the mean pairwise correlation between structures within families of log read coverage ($\log_2$(reads/base+1))(filtered GW set; novel elements only), across the 16 tissues of a large RNA-seq dataset (Illumina Inc. 2010) was compared with the mean correlation distribution of random families generated by shuffling of members between families. 1000 random shuffles were used. Low expression (<2 reads/base) structures were set to 0 to limit noise. Then structures showing no expression in any tissue were excluded, as were structures showing strong homology with the mitochondrial chromosome, which may represent pseudogenes, as well as duplicate structures repeated in the same gene (a single representative member was randomly chosen in this case). Then the

mean pair-wise Pearson correlation coefficient was computed per family, and the overall mean (weighted by family size) was computed. P-values were computed from this permutation test.

*Expression correlation analysis of POP1 structure:*

Affymetrix Human Exon 1.0 array probes overlapping the region of the *POP1* intronic structure were used for a correlative expression analysis across 11 human tissues (Pohl et al. 2009)). The mean of the probe expression overlying the *POP1* structure and 5′ adjacent region (consisting of probe ids:3108683, 3108684, 3108685) were compared with a probe overlying an upstream exon (probe id:3108681). Pearson correlation coefficient showed substantial negative correlation (Pearson correlation coefficient -0.63; p-value=0.018 (one-sided test)). (Note: probe 3108684 is uniquely mapped by BLAT; probe 3108683 also maps to a region of a low expressed non-tRNA gene, so cross-hybridization is not a likely cause of the negative correlation noted).

## References:

Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE et al. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**(7146): 799-816.

Carey V, Long L, Gentleman R. 2010. RBGL: An interface to the BOOST graph library.

Christensen DE, Brzovic PS, Klevit RE. 2007. E2-BRCA1 RING interactions dictate synthesis of mono- or specific polyubiquitin chain linkages. *Nat Struct Mol Biol* **14**(10): 941-948.

Ciaudo C, Bourdet A, Cohen-Tannoudji M, Dietz HC, Rougeulle C, Avner P. 2006. Nuclear mRNA degradation pathway(s) are implicated in Xist regulation and X chromosome inactivation. *PLoS Genet* **2**(6): e94.

Durbin R. 1998. *Biological sequence analysis : probabalistic models of proteins and nucleic acids*. Cambridge University Press, Cambridge, UK New York.

Eddy SR, Durbin R. 1994. RNA sequence analysis using covariance models. *Nucleic Acids Res* **22**(11): 2079-2088.

Eloranta TO. 1977. Tissue distribution of S-adenosylmethionine and S-adenosylhomocysteine in the rat. Effect of age, sex and methionine administration on the metabolism of S-adenosylmethionine, S-adenosylhomocysteine and polyamines. *Biochem J* **166**(3): 521-529.

Faustino NA, Cooper TA. 2003. Pre-mRNA splicing and human disease. *Genes Dev* **17**(4): 419-437.

Gesell T, Washietl S. 2008. Dinucleotide controlled null models for comparative RNA gene prediction. *BMC Bioinformatics* **9**: 248.

Gruber AR, Findeiss S, Washietl S, Hofacker IL, Stadler PF. 2010. Rnaz 2.0: Improved Noncoding Rna Detection. *Pac Symp Biocomput* **15**: 69-79.

Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP et al. 2009. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**(7235): 223-227.

Hall LL, Byron M, Pageau G, Lawrence JB. 2009. AURKB-mediated effects on chromatin regulate binding versus release of XIST RNA to the inactive chromosome. *J Cell Biol* **186**(4): 491-507.

Hartuv E, Shamir R. 2000. A clustering algorithm based on graph connectivity. *Information Processing Letters* **76**(4-6): 175-181

Holt CE, Bullock SL. 2009. Subcellular mRNA localization in animal cells and why it matters. *Science* **326**(5957): 1212-1216.

Hussain S, Zhang Y, Galardy PJ. 2009. DUBs and cancer: the role of deubiquitinating enzymes as oncogenes, non-oncogenes and tumor suppressors. *Cell Cycle* **8**(11): 1688-1697.

Illumina Inc. 2010. Body Map 2.0 RNA-seq dataset. San Diego, California.

Juang BH, Rabiner LR. 1985. A probabilistic distance measure for hidden Markov models. *AT&T Tech J* **64**(2): 391-408.

Karro JE, Yan Y, Zheng D, Zhang Z, Carriero N, Cayting P, Harrrison P, Gerstein M. 2007. Pseudogene.org: a comprehensive database and comparison platform for pseudogene annotation. *Nucleic Acids Res* **35**(Database issue): D55-60.

Kent WJ. 2002. BLAT--the BLAST-like alignment tool. *Genome Res* **12**(4): 656-664.

Khalil AM, Guttman M, Huarte M, Garber M, Raj A, Rivea Morales D, Thomas K, Presser A, Bernstein BE, van Oudenaarden A et al. 2009. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc Natl Acad Sci U S A* **106**(28): 11667-11672.

Koeller DM, Casey JL, Hentze MW, Gerhardt EM, Chan LN, Klausner RD, Harford JB. 1989. A cytosolic protein binds to structural elements within the iron regulatory region of the transferrin receptor mRNA. *Proc Natl Acad Sci U S A* **86**(10): 3574-3578.

Lampson MA, Renduchitala K, Khodjakov A, Kapoor TM. 2004. Correcting improper chromosome-spindle attachments during cell division. *Nat Cell Biol* **6**(3): 232-237.

Lomeli H, Mosbacher J, Melcher T, Hoger T, Geiger JR, Kuner T, Monyer H, Higuchi M, Bach A, Seeburg PH. 1994. Control of kinetic properties of AMPA receptor channels by nuclear RNA editing. *Science* **266**(5191): 1709-1713.

Maenner S, Blaud M, Fouillen L, Savoye A, Marchand V, Dubois A, Sanglier-Cianferani S, Van Dorsselaer A, Clerc P, Avner P et al. 2010. 2-D structure of the A region of Xist RNA and its implication for PRC2 association. *PLoS Biol* **8**(1): e1000276.

Mayr C, Bartel DP. 2009. Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell* **138**(4): 673-684.

McAlinden A, Havlioglu N, Liang L, Davies SR, Sandell LJ. 2005. Alternative splicing of type II procollagen exon 2 is regulated by the combination of a weak 5' splice site and an adjacent intronic stem-loop cis element. *J Biol Chem* **280**(38): 32700-32711.

Nakaya HI, Amaral PP, Louro R, Lopes A, Fachel AA, Moreira YB, El-Jundi TA, da Silva AM, Reis EM, Verjovski-Almeida S. 2007. Genome mapping and expression analyses of human intronic noncoding RNAs reveal tissue-specific patterns and enrichment in genes related to regulation of transcription. *Genome Biol* **8**(3): R43.

Nishikawa H, Wu W, Koike A, Kojima R, Gomi H, Fukuda M, Ohta T. 2009. BRCA1-associated protein 1 interferes with BRCA1/BARD1 RING heterodimer activity. *Cancer Res* **69**(1): 111-119.

Park SW, Kang Y, Sypula JG, Choi J, Oh H, Park Y. 2007. An evolutionarily conserved domain of roX2 RNA is sufficient for induction of H4-Lys16 acetylation on the Drosophila X chromosome. *Genetics* **177**(3): 1429-1437.

Pedersen JS, Bejerano G, Siepel A, Rosenbloom K, Lindblad-Toh K, Lander ES, Kent J, Miller W, Haussler D. 2006. Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput Biol* **2**(4): e33.

Pohl AA, Sugnet CW, Clark TA, Smith K, Fujita PA, Cline MS. 2009. Affy exon tissues: exon levels in normal tissues in human, mouse and rat. *Bioinformatics* **25**(18): 2442-2443.

Pollard KS, Salama SR, Lambert N, Lambot MA, Coppens S, Pedersen JS, Katzman S, King B, Onodera C, Siepel A et al. 2006. An RNA gene expressed during cortical development evolved rapidly in humans. *Nature* **443**(7108): 167-172.

Pupko T, Pe'er I, Shamir R, Graur D. 2000. A fast algorithm for joint reconstruction of ancestral amino acid sequences. *Mol Biol Evol* **17**(6): 890-896.

Satterwhite E, Sonoki T, Willis TG, Harder L, Nowak R, Arriola EL, Liu H, Price HP, Gesk S, Steinemann D et al. 2001. The BCL11 gene family: involvement of BCL11A in lymphoid malignancies. *Blood* **98**(12): 3413-3420.

Shirahata E, Iwasaki H, Takagi M, Lin C, Bennett V, Okamura Y, Hayasaka K. 2006. Ankyrin-G regulates inactivation gating of the neuronal sodium channel, Nav1.6. *J Neurophysiol* **96**(3): 1347-1357.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**(8): 1034-1050.

Stark A, Lin MF, Kheradpour P, Pedersen JS, Parts L, Carlson JW, Crosby MA, Rasmussen MD, Roy S, Deoras AN et al. 2007. Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. *Nature* **450**(7167): 219-232.

Stefanovic B, Brenner DA. 2003. 5' stem-loop of collagen alpha 1(I) mRNA inhibits translation in vitro but is required for triple helical collagen synthesis in vivo. *J Biol Chem* **278**(2): 927-933.

Sunwoo H, Dinger ME, Wilusz JE, Amaral PP, Mattick JS, Spector DL. 2009. MEN epsilon/beta nuclear-retained non-coding RNAs are up-regulated upon muscle differentiation and are essential components of paraspeckles. *Genome Res* **19**(3): 347-359.

Trockenbacher A, Suckow V, Foerster J, Winter J, Krauss S, Ropers HH, Schneider R, Schweiger S. 2001. MID1, mutated in Opitz syndrome, encodes an ubiquitin ligase that targets phosphatase 2A for degradation. *Nat Genet* **29**(3): 287-294.

Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**(7221): 470-476.

Washietl S, Hofacker IL. 2004. Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics. *J Mol Biol* **342**(1): 19-30.

Washietl S, Hofacker IL, Stadler PF. 2005. Fast and reliable prediction of noncoding RNAs. *Proc Natl Acad Sci U S A* **102**(7): 2454-2459.

Wilusz JE, Freier SM, Spector DL. 2008. 3' end processing of a long nuclear-retained noncoding RNA yields a tRNA-like cytoplasmic RNA. *Cell* **135**(5): 919-932.

Wilusz JE, Spector DL. 2010. An unexpected ending: noncanonical 3' end processing mechanisms. *RNA* **16**(2): 259-266.

Xu N, Donohoe ME, Silva SS, Lee JT. 2007. Evidence that homologous X-chromosome pairing requires transcription and Ctcf protein. *Nat Genet* **39**(11): 1390-1396.

Xu N, Tsai CL, Lee JT. 2006. Transient homologous chromosome pairing marks the onset of X inactivation. *Science* **311**(5764): 1149-1152.

Yang Y, Kitagaki J, Wang H, Hou DX, Perantoni AO. 2009. Targeting the ubiquitin-proteasome system for cancer therapy. *Cancer Sci* **100**(1): 24-28.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**(8): 1586-1591.