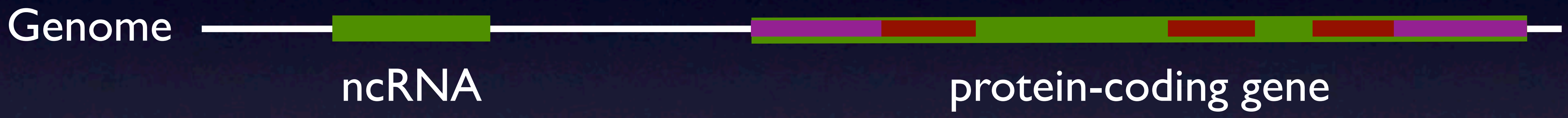
Comparative identification of novel human structural RNA families

Jakob Skou Pedersen
Department of Molecular Medicine
Aarhus University, Denmark

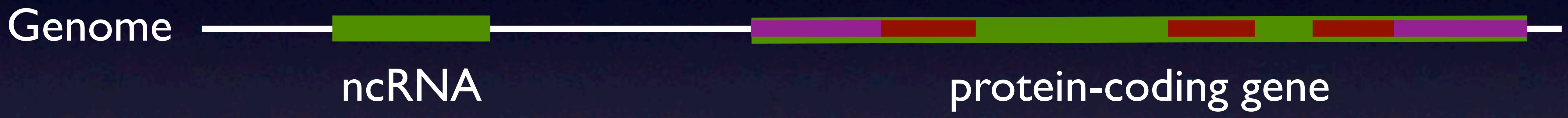
Benasque 2012

Structural RNA



Definition: Any RNA sequence that folds into a structure of functional importance

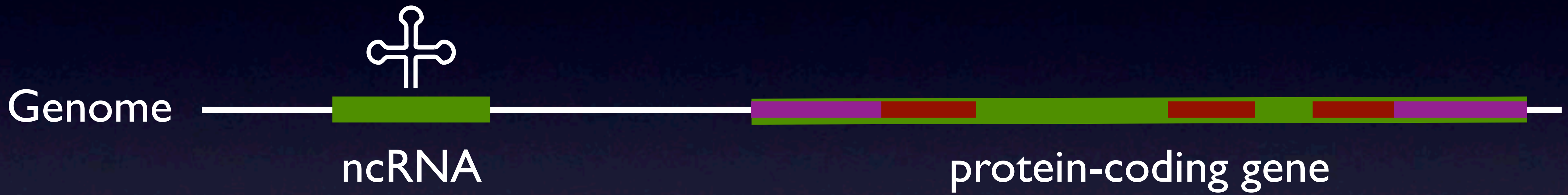
Structural RNA



Definition: Any RNA sequence that folds into a structure of functional importance

Such as:

Structural RNA

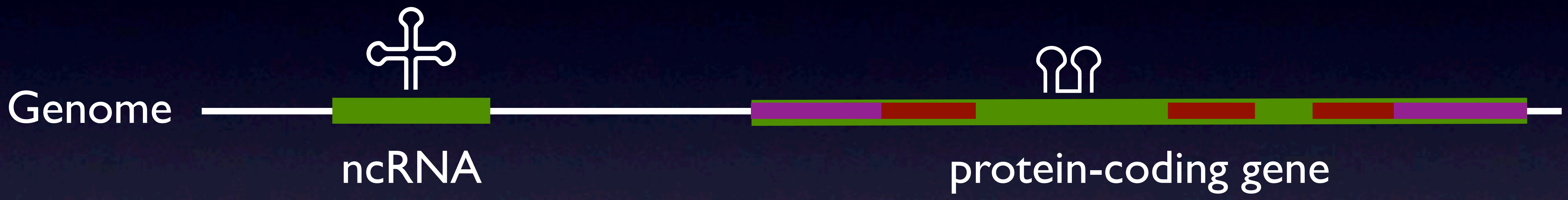


Definition: Any RNA sequence that folds into a structure of functional importance

Such as:

- independently transcribed ncRNAs

Structural RNA

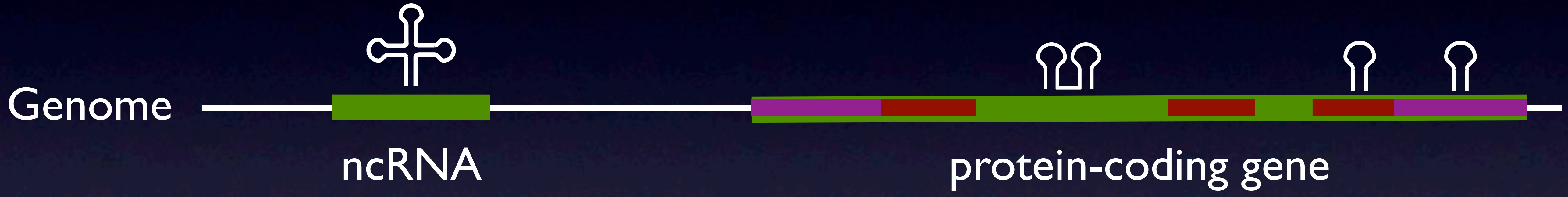


Definition: Any RNA sequence that folds into a structure of functional importance

Such as:

- independently transcribed ncRNAs
- ncRNAs excised from longer transcripts

Structural RNA

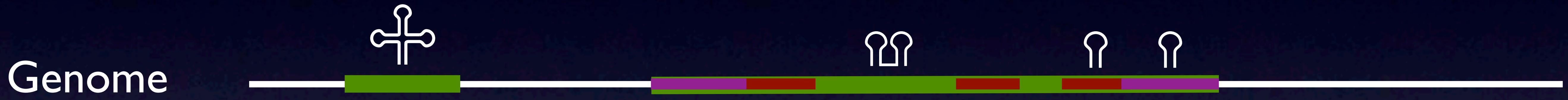


Definition: Any RNA sequence that folds into a structure of functional importance

Such as:

- independently transcribed ncRNAs
- ncRNAs excised from longer transcripts
- cis-regulatory elements within protein-coding genes and ncRNAs

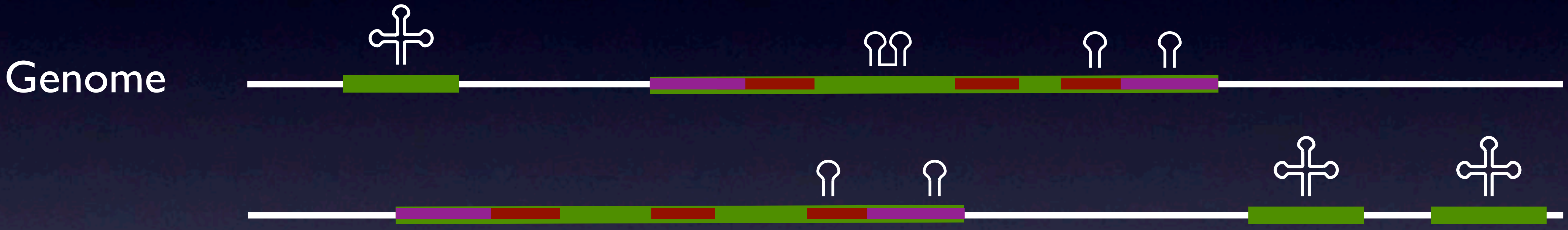
Families of structural RNA



Family members share ancestry

- Duplications may be local or far apart

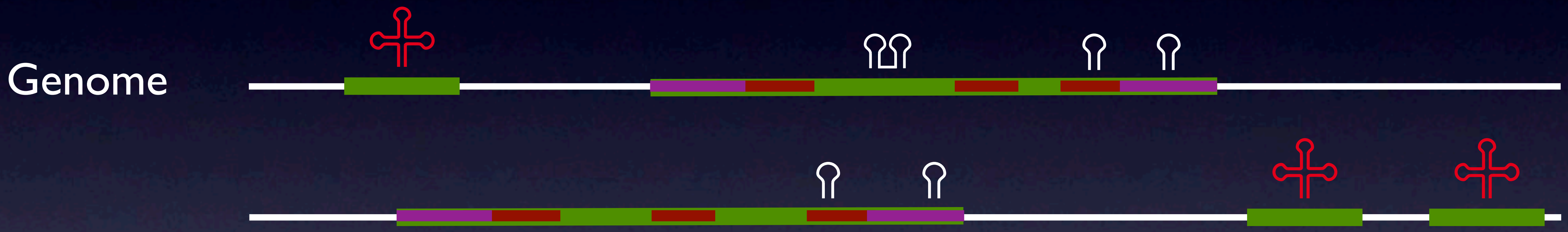
Families of structural RNA



Family members share ancestry

- Duplications may be local or far apart

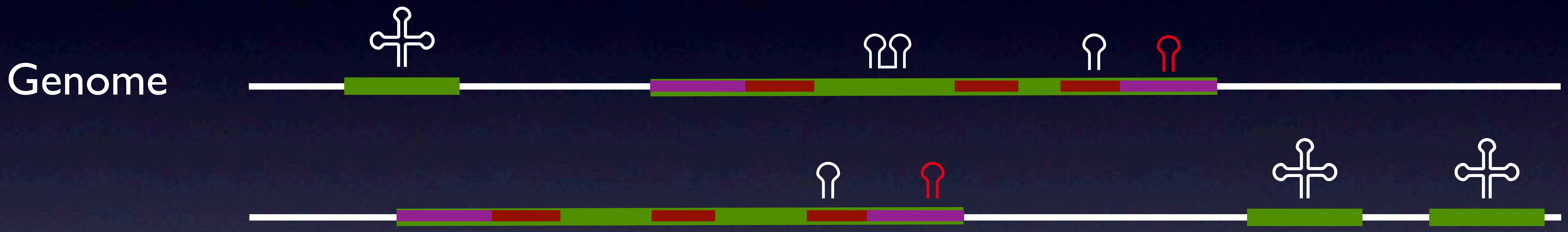
Families of structural RNA



Family members share ancestry

- Duplications may be local or far apart

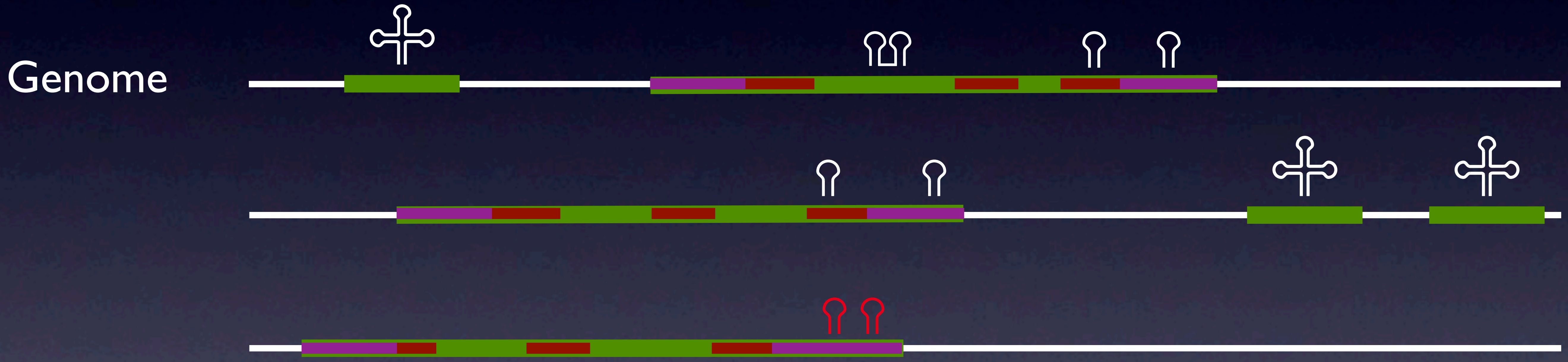
Families of structural RNA



Family members share ancestry

- Duplications may be local or far apart
- Cis-regulatory families often reflect protein-coding gene families

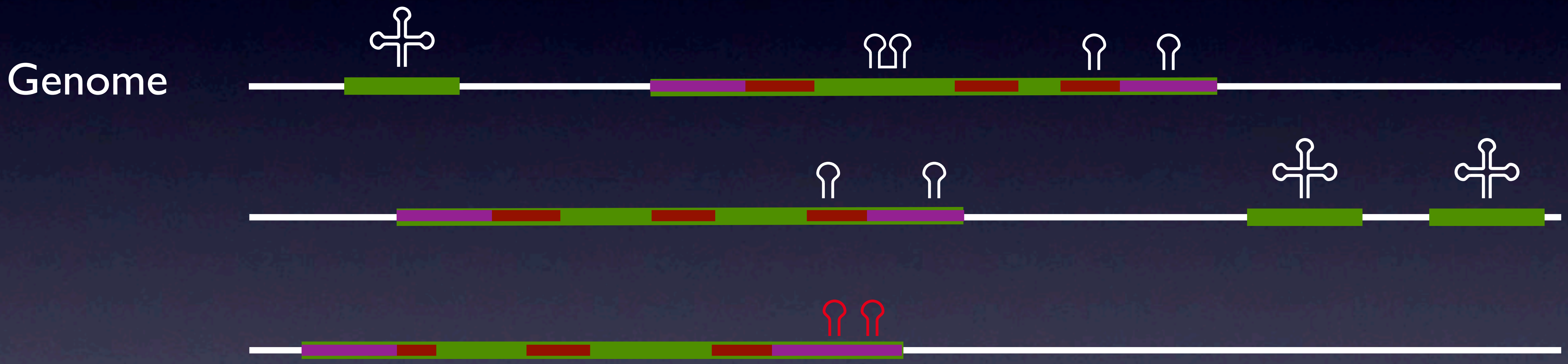
Families of structural RNA



Family members share ancestry

- Duplications may be local or far apart
- Cis-regulatory families often reflect protein-coding gene families

Families of structural RNA



Family members share ancestry

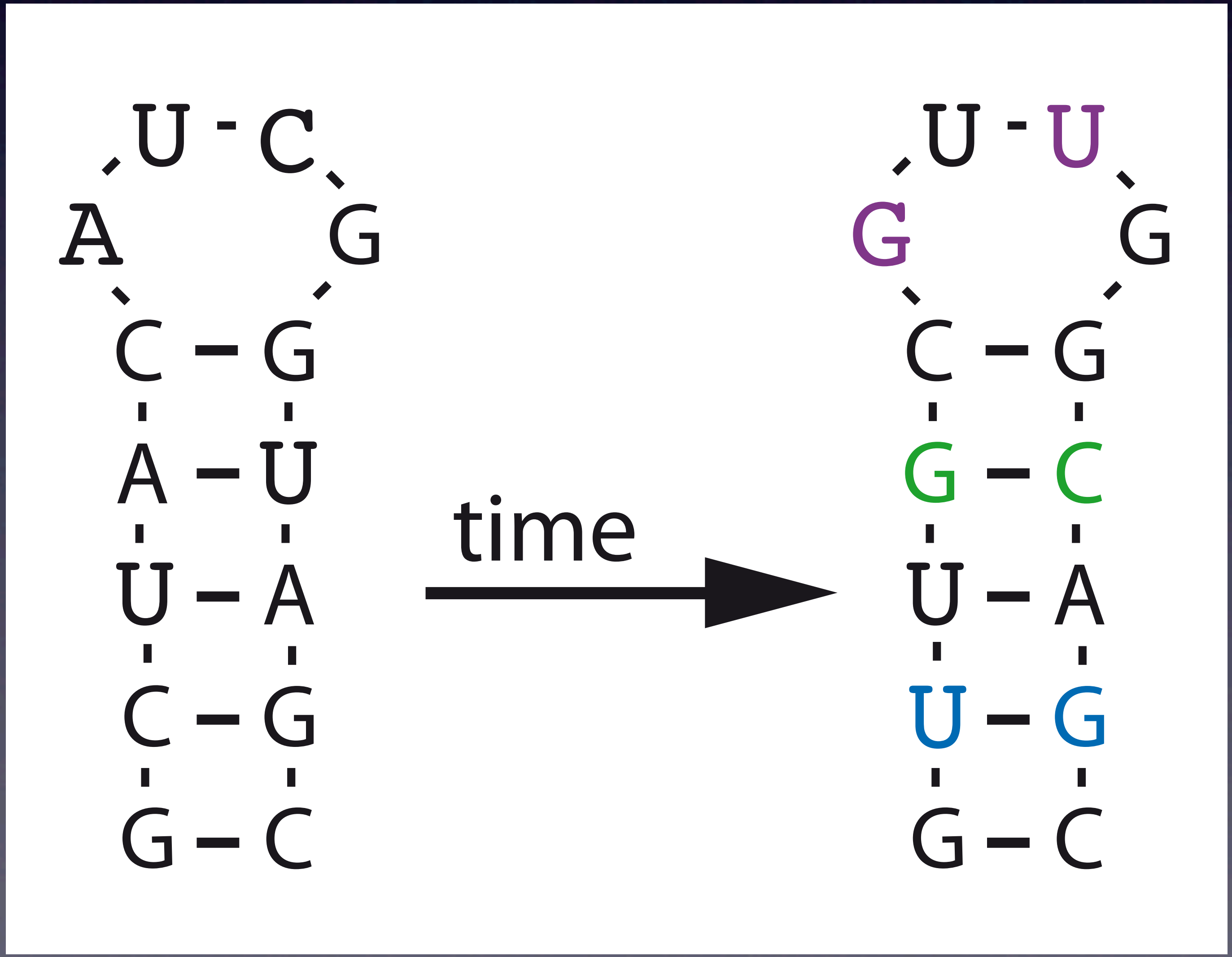
- Duplications may be local or far apart
- Cis-regulatory families often reflect protein-coding gene families

For simple structures convergent evolution may be possible

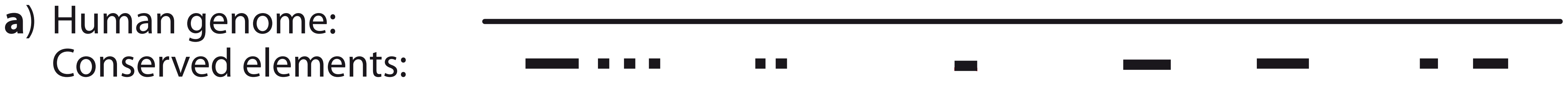
Genomic structure screen

Evolution constrained by structure

Characteristic substitution pattern



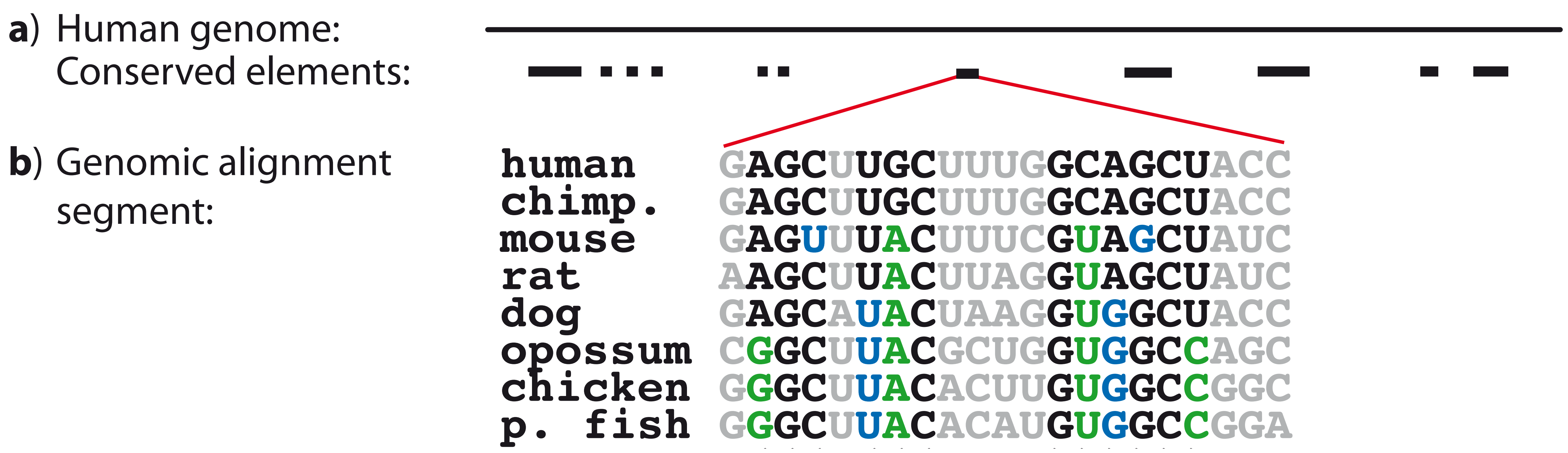
EvoFold structure prediction



a) Human genome:
Conserved elements:

Pedersen et al., 2006, PLoS Computational Biology.
Knudsen & Hein, 1999, Bioinformatics.

EvoFold structure prediction



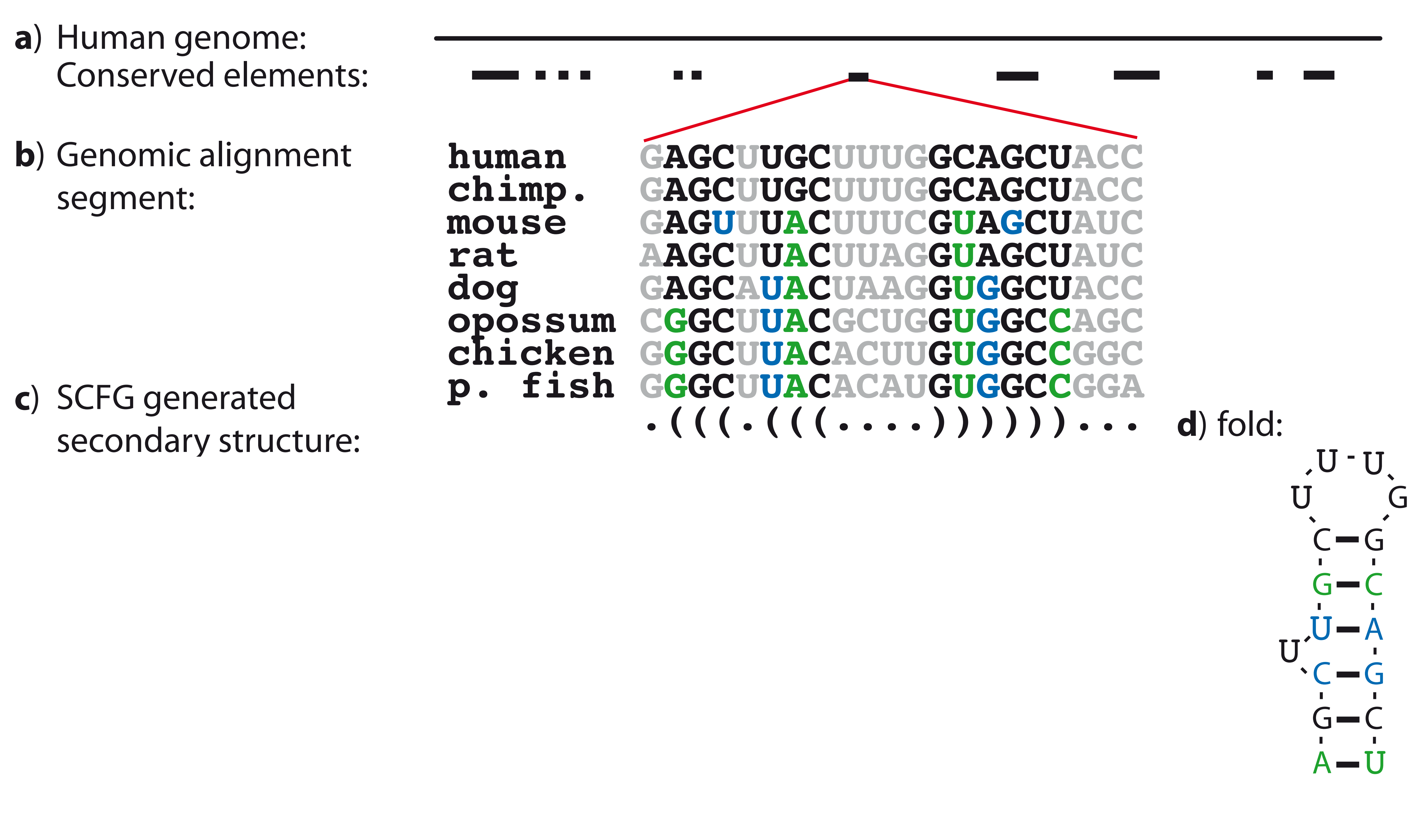
a) Human genome:
Conserved elements:

b) Genomic alignment segment:

```
human    GAGCUUGCUUUGGCAGCUACC
chimp.   GAGCUUGCUUUGGCAGCUACC
mouse    GAGUUUACUUUCGUAGCUAUC
rat      AAGCUUACUUAGGUAGCUAUC
dog      GAGCAUACUAAGGUGGCUACC
opossum  CGGCUUACGCUGGUGGCCAGC
chicken  GGGCUUACACUUGUGGCCGGC
p. fish  GGGCUUACACAUGUGGCCGGA
```

Pedersen et al., 2006, PLoS Computational Biology.
Knudsen & Hein, 1999, Bioinformatics.

EvoFold structure prediction



a) Human genome:
Conserved elements:

b) Genomic alignment segment:

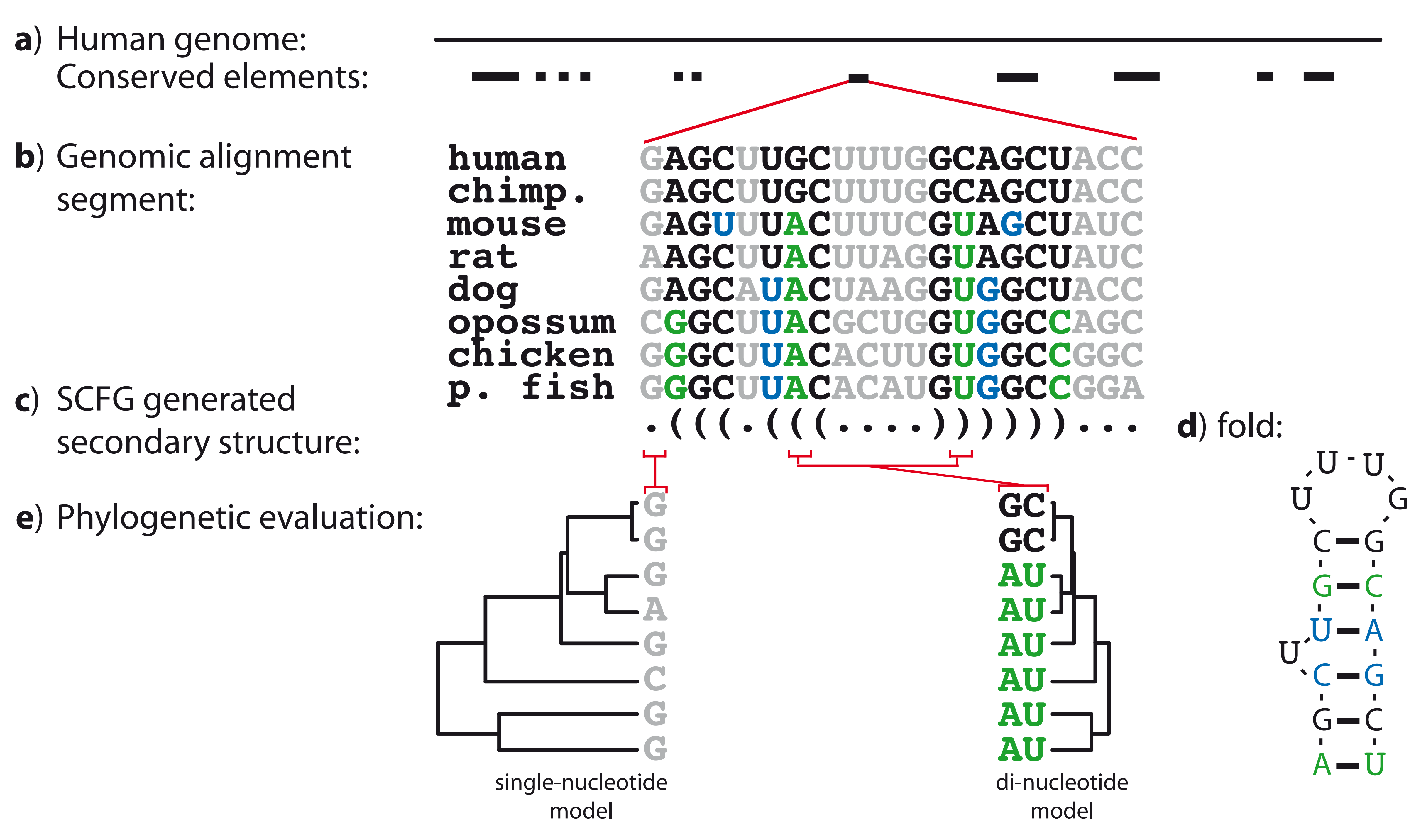
```
human    GAGCUUGCUUUGGCAGCUACC
chimp.   GAGCUUGCUUUGGCAGCUACC
mouse    GAGUUUACUUUCGUAGCUAUC
rat      AAGCUUACUUAGGUAGCUAUC
dog      GAGCAUACUAAGGUGGCUACC
opossum  CGGCUUACGCUGGUGGCCAGC
chicken  GGGCUUACACUUGUGGCCGGC
p. fish  GGGCUUACACAUGUGGCCGGA
         .(((.(((....))))))...
```

c) SCFG generated secondary structure:

d) fold:

Pedersen et al., 2006, PLoS Computational Biology.
Knudsen & Hein, 1999, Bioinformatics.

EvoFold structure prediction



a) Human genome:
Conserved elements:

b) Genomic alignment segment:

```
human    GAGCUUGCUUUGGCAGCUACC
chimp.   GAGCUUGCUUUGGCAGCUACC
mouse    GAGUUUACUUUCGUAGCUAUC
rat      AAGCUUACUUAGGUAGCUAUC
dog      GAGCAUACUAAGGUGGCUACC
opossum  CGGCUUACGCUGGUGGCCAGC
chicken  GGGCUUACACUUGUGGCCGGC
p. fish  GGGCUUACACAUGUGGCCGGA
         .(((.(((....))))))...
```

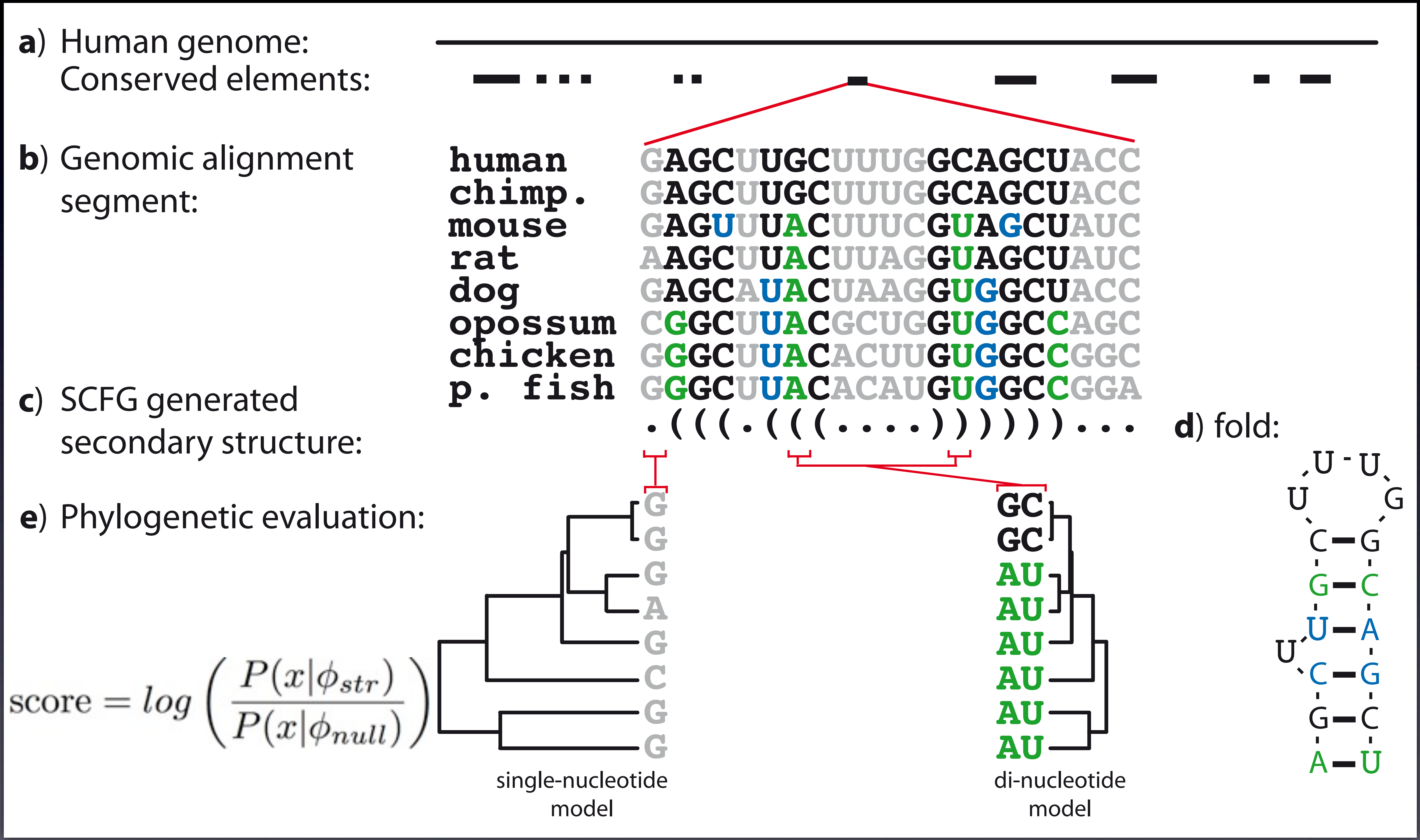
c) SCFG generated secondary structure:

d) fold:

e) Phylogenetic evaluation:

Pedersen et al., 2006, PLoS Computational Biology.
Knudsen & Hein, 1999, Bioinformatics.

EvoFold structure prediction



a) Human genome:
Conserved elements:

b) Genomic alignment segment:

```
human    GAGCUUGCUUUGGCAGCUACC
chimp.   GAGCUUGCUUUGGCAGCUACC
mouse    GAGUUUACUUUCGUAGCUAUC
rat      AAGCUUACUUAGGUAGCUAUC
dog      GAGCAUACUAAGGUGGCUACC
opossum  CGGCUUACGCUGGUGGCCAGC
chicken  GGGCUUACACUUGUGGCCGGC
p. fish  GGGCUUACACAUGUGGCCGGA
         .(((.((((....)))))))...
```

c) SCFG generated secondary structure:

d) fold:

e) Phylogenetic evaluation:

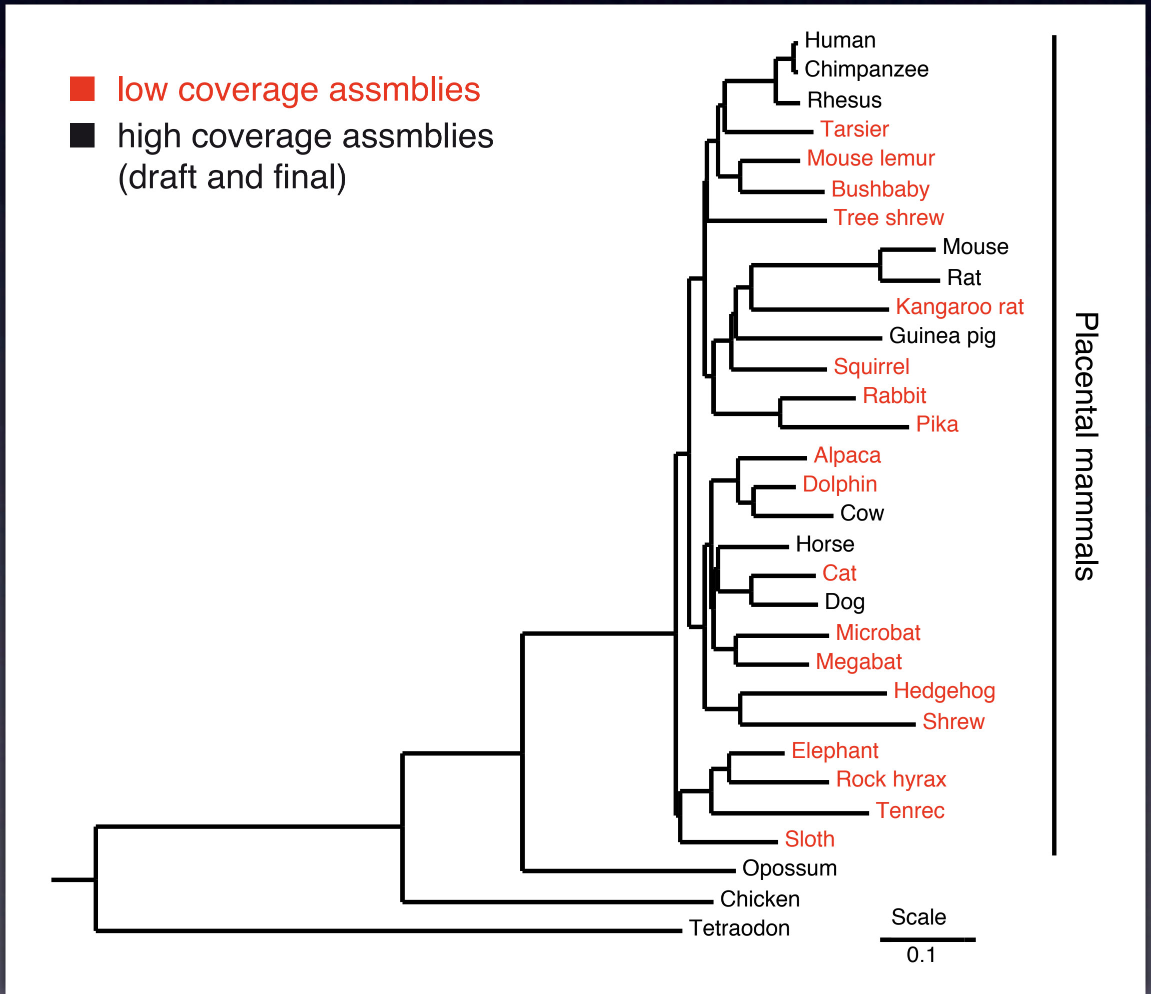
$$\text{score} = log\left(\frac{P(x|\phi_{str})}{P(x|\phi_{null})}\right)$$

Pedersen et al., 2006, PLoS Computational Biology.
Knudsen & Hein, 1999, Bioinformatics.

Screen of 31-way vertebrate alignments

Input: conserved alignment segments (5.6% of genome)

Phylogenetic tree of input species



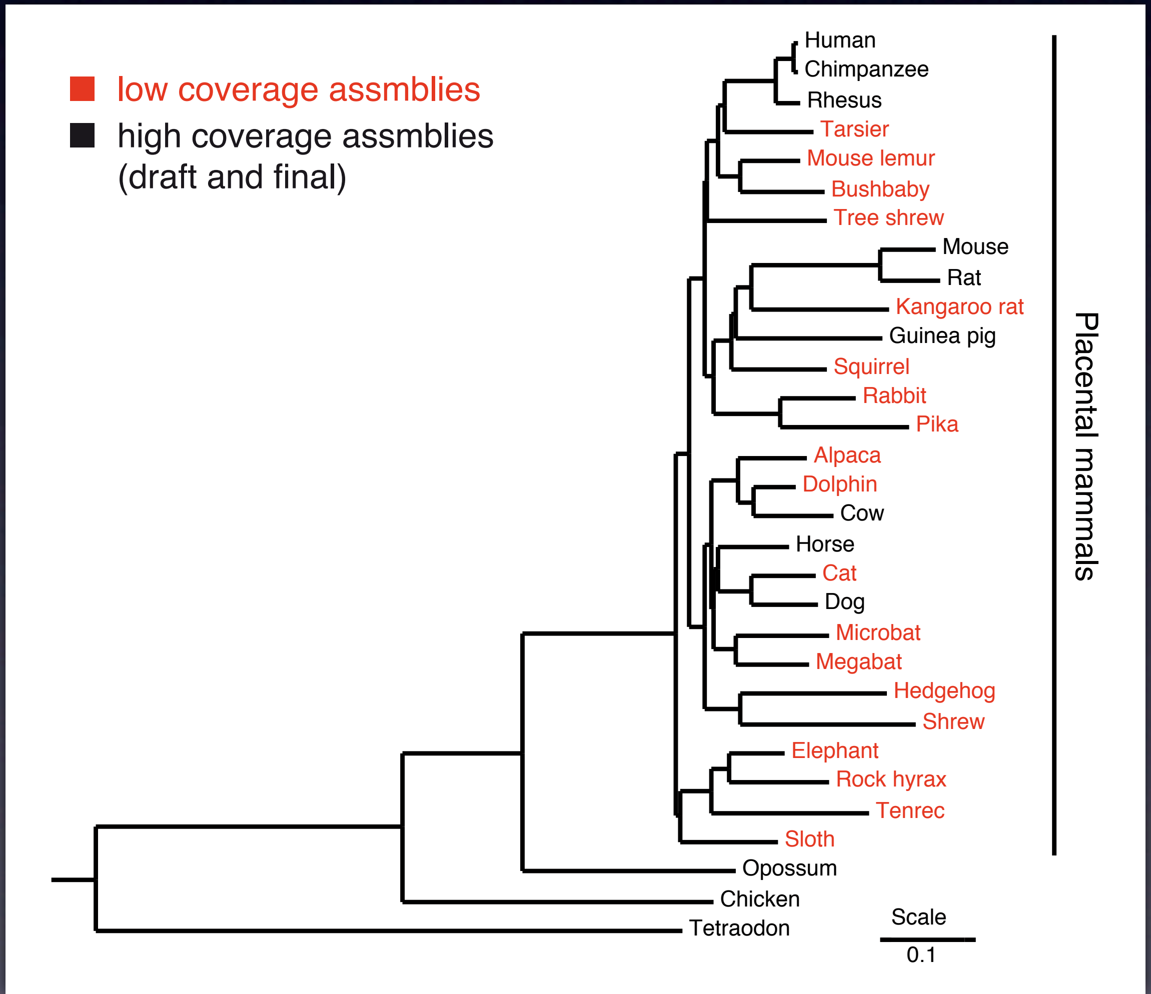
Data from 29 mammals sequencing and analysis consortium: Kerstin Lindblad-Toh. A high-resolution map of human evolutionary constraint using 29 mammals. Nature (2011).
Brian J. Parker, et al. New families of human regulatory RNA structures identified by comparative analysis of vertebrate genomes. Genome Research (2011).

Screen of 31-way vertebrate alignments

Input: conserved alignment segments (5.6% of genome)

Output: 37,381 predictions (0.05% of genome)

Phylogenetic tree of input species



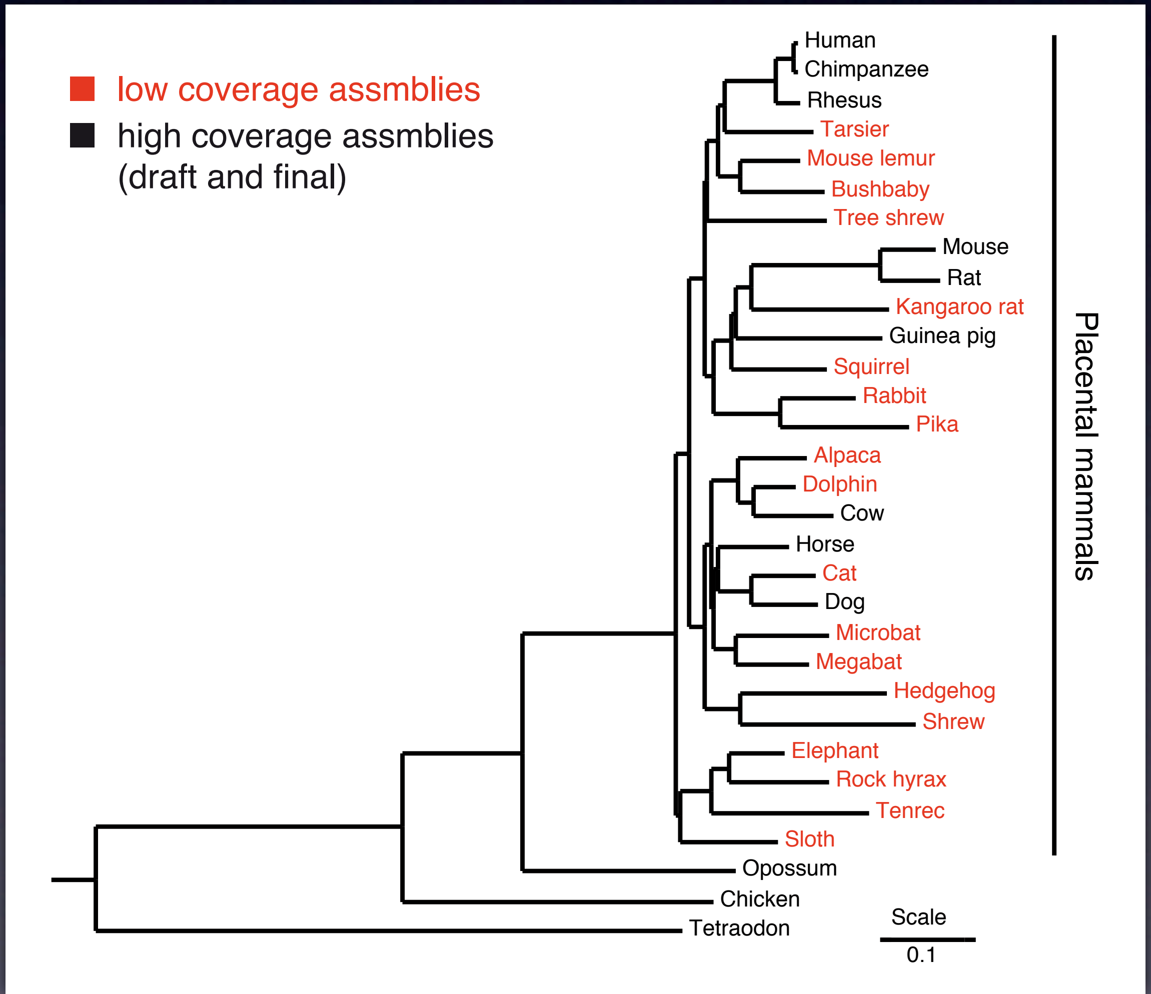
Data from 29 mammals sequencing and analysis consortium: Kerstin Lindblad-Toh. A high-resolution map of human evolutionary constraint using 29 mammals. Nature (2011).
Brian J. Parker, et al. New families of human regulatory RNA structures identified by comparative analysis of vertebrate genomes. Genome Research (2011).

Screen of 31-way vertebrate alignments

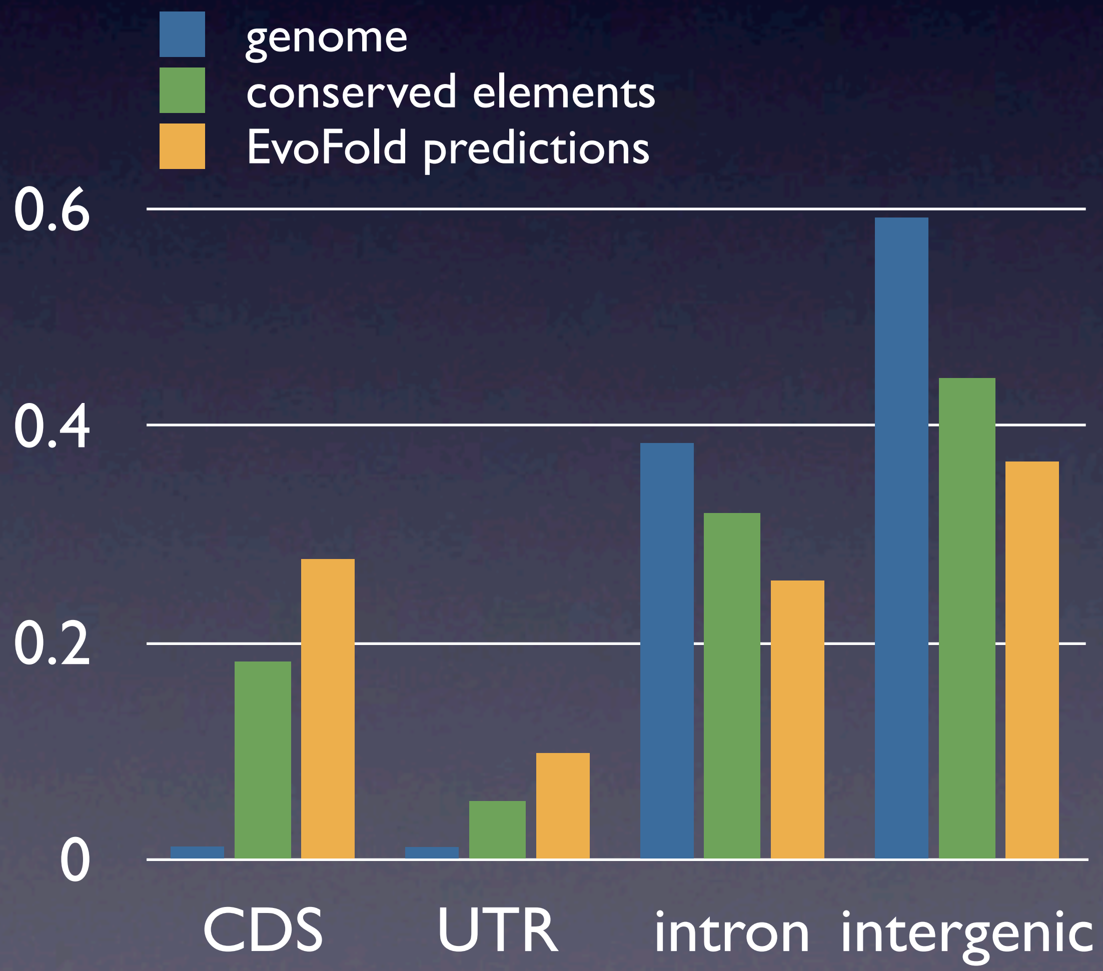
Input: conserved alignment segments (5.6% of genome)

Output: 37,381 predictions (0.05% of genome)

Phylogenetic tree of input species



Genomic distribution



Data from 29 mammals sequencing and analysis consortium: Kerstin Lindblad-Toh. A high-resolution map of human evolutionary constraint using 29 mammals. Nature (2011).

Brian J. Parker, et al. New families of human regulatory RNA structures identified by comparative analysis of vertebrate genomes. Genome Research (2011).

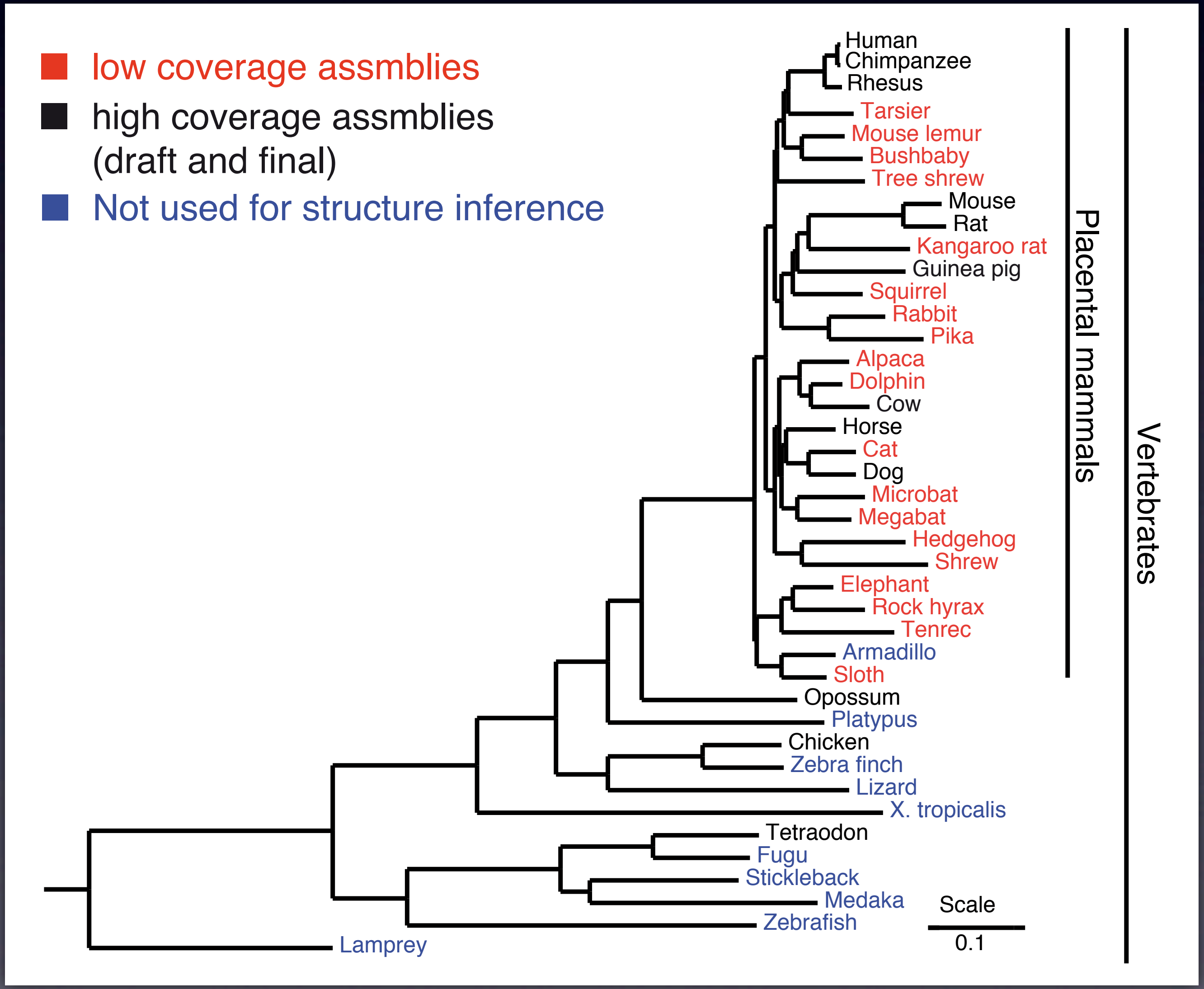
Significance evaluation using additional vertebrate genomes

EvoP method

Question: how surprising is the observed number of double substitutions?

- Monte Carlo approach:
  - Simulate iid substitutions across columns on phylogeny.
  - Count double substitutions given structure.
  - Estimate P-value as fraction simulations with at least as many double substitutions.

Ten vertebrates genomes not used for structure inference



Brian J. Parker, et al. New families of human regulatory RNA structures identified by comparative analysis of vertebrate genomes. Genome Research (2011).

Significance evaluation using additional vertebrate genomes

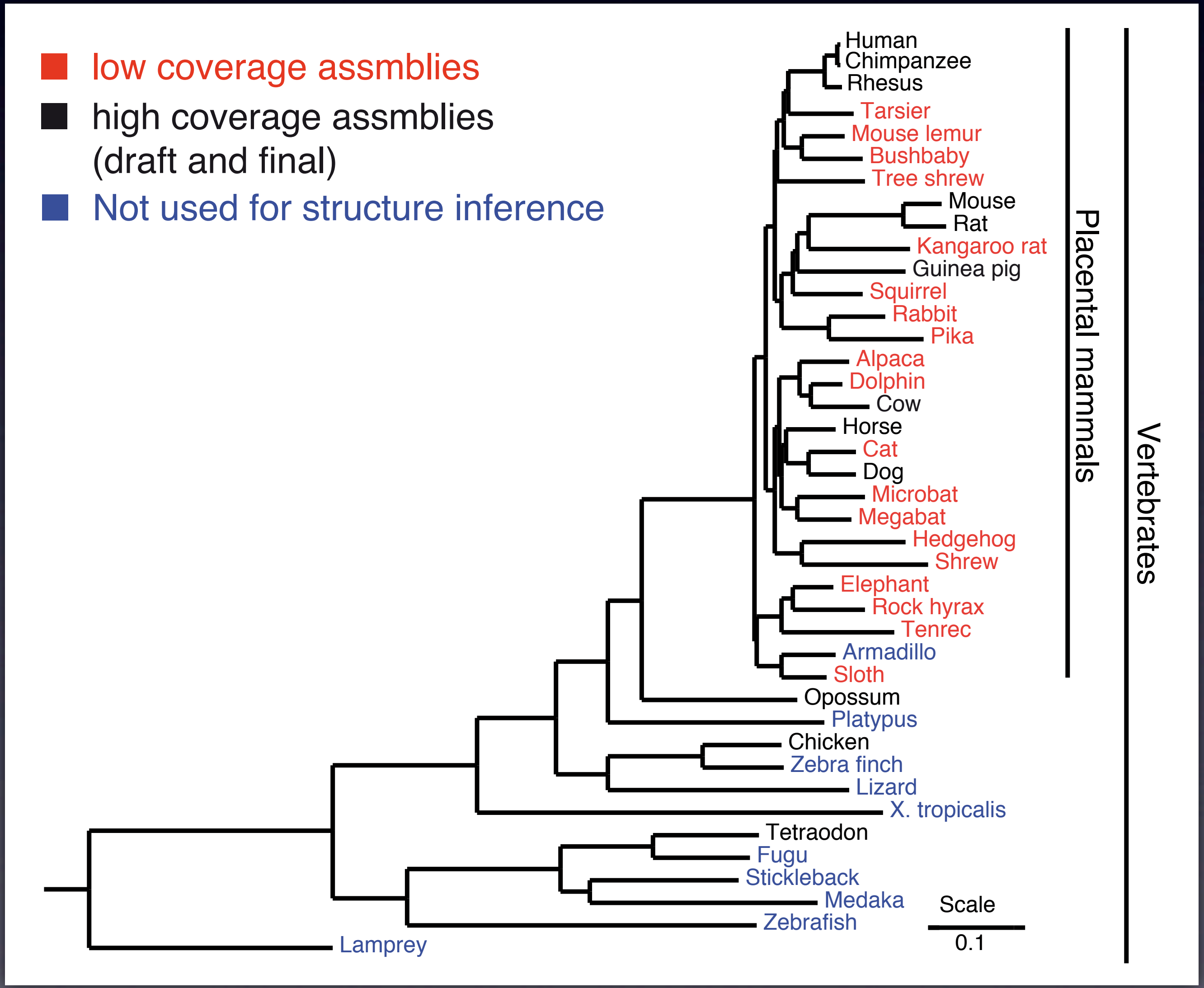
EvoP method

Question: how surprising is the observed number of double substitutions?

- Monte Carlo approach:
  - Simulate iid substitutions across columns on phylogeny.
  - Count double substitutions given structure.
  - Estimate P-value as fraction simulations with at least as many double substitutions.

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

Ten vertebrates genomes not used for structure inference



Brian J. Parker, et al. New families of human regulatory RNA structures identified by comparative analysis of vertebrate genomes. Genome Research (2011).

Significance evaluation using additional vertebrate genomes

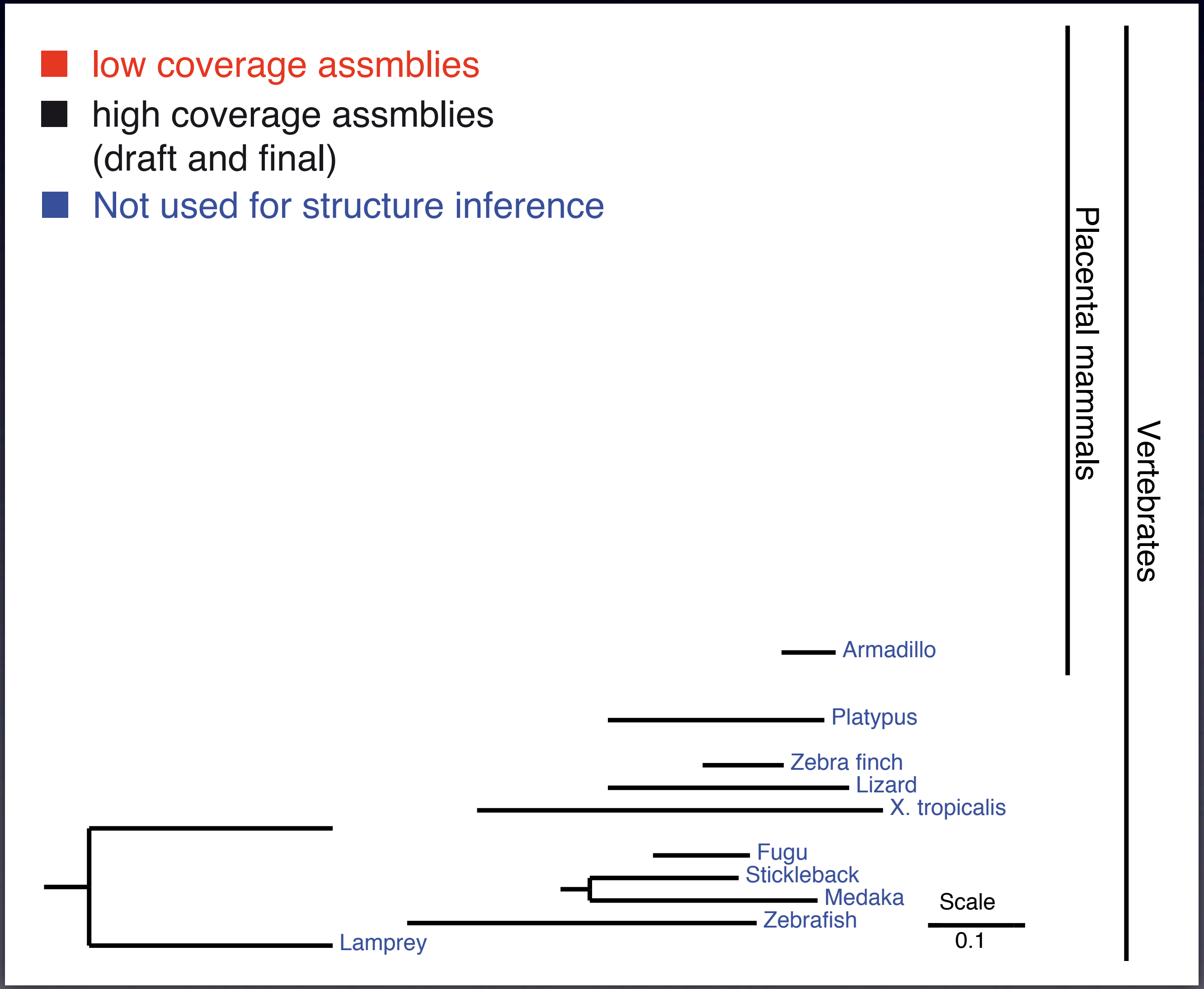
EvoP method

Question: how surprising is the observed number of double substitutions?

- Monte Carlo approach:
  - Simulate iid substitutions across columns on phylogeny.
  - Count double substitutions given structure.
  - Estimate P-value as fraction simulations with at least as many double substitutions.

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

Ten vertebrates genomes not used for structure inference



Brian J. Parker, et al. New families of human regulatory RNA structures identified by comparative analysis of vertebrate genomes. Genome Research (2011).

Significance evaluation using additional vertebrate genomes

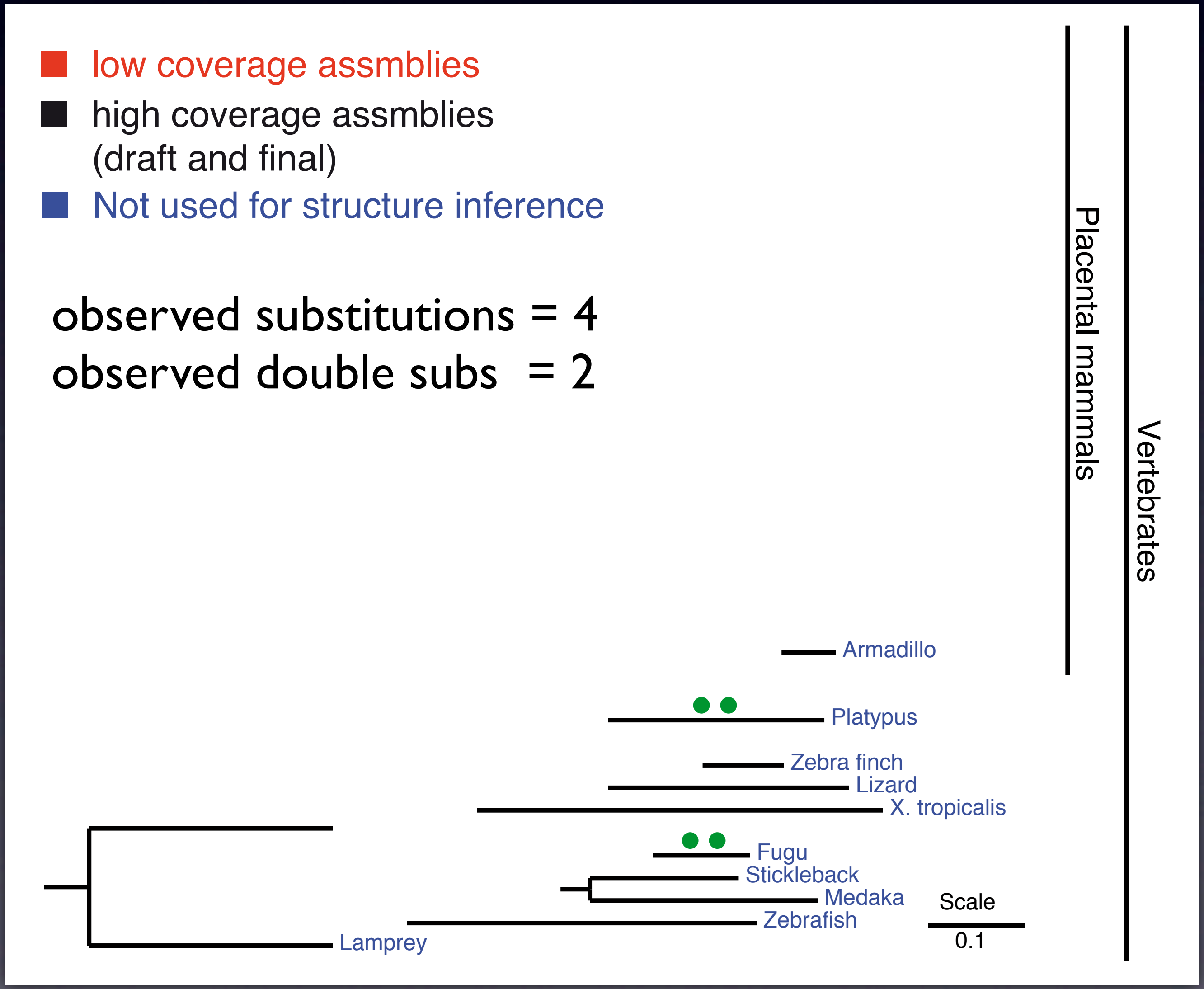
EvoP method

Question: how surprising is the observed number of double substitutions?

- Monte Carlo approach:
  - Simulate iid substitutions across columns on phylogeny.
  - Count double substitutions given structure.
  - Estimate P-value as fraction simulations with at least as many double substitutions.

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

Ten vertebrates genomes not used for structure inference



🟥 low coverage assmblies
⬛ high coverage assmblies (draft and final)
🟦 Not used for structure inference

observed substitutions = 4
observed double subs = 2

Brian J. Parker, et al. New families of human regulatory RNA structures identified by comparative analysis of vertebrate genomes. Genome Research (2011).

Significance evaluation using additional vertebrate genomes

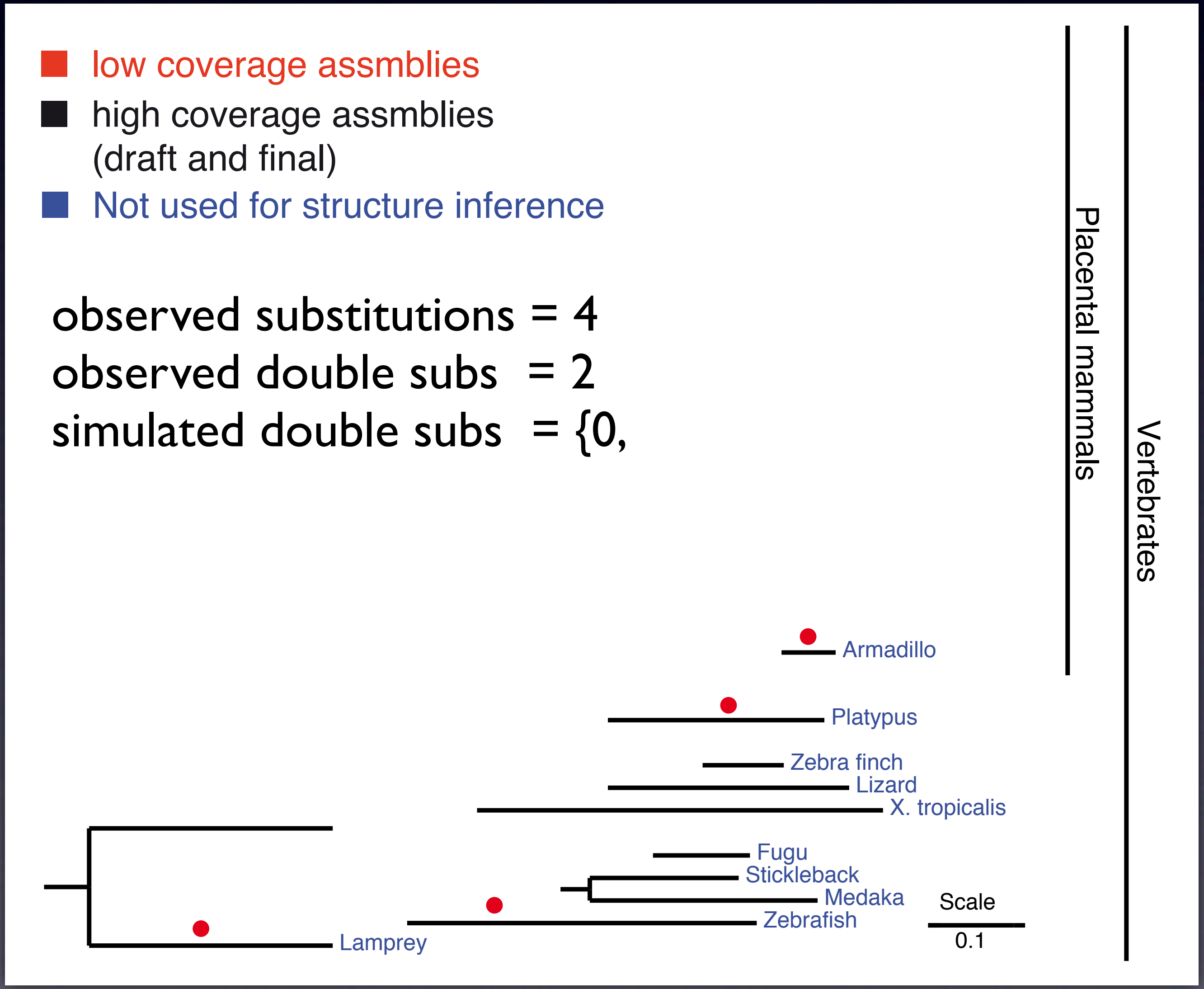
EvoP method

Question: how surprising is the observed number of double substitutions?

- Monte Carlo approach:
  - Simulate iid substitutions across columns on phylogeny.
  - Count double substitutions given structure.
  - Estimate P-value as fraction simulations with at least as many double substitutions.

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

Ten vertebrates genomes not used for structure inference



observed substitutions = 4
observed double subs = 2
simulated double subs = {0,

Brian J. Parker, et al. New families of human regulatory RNA structures identified by comparative analysis of vertebrate genomes. Genome Research (2011).

Significance evaluation using additional vertebrate genomes

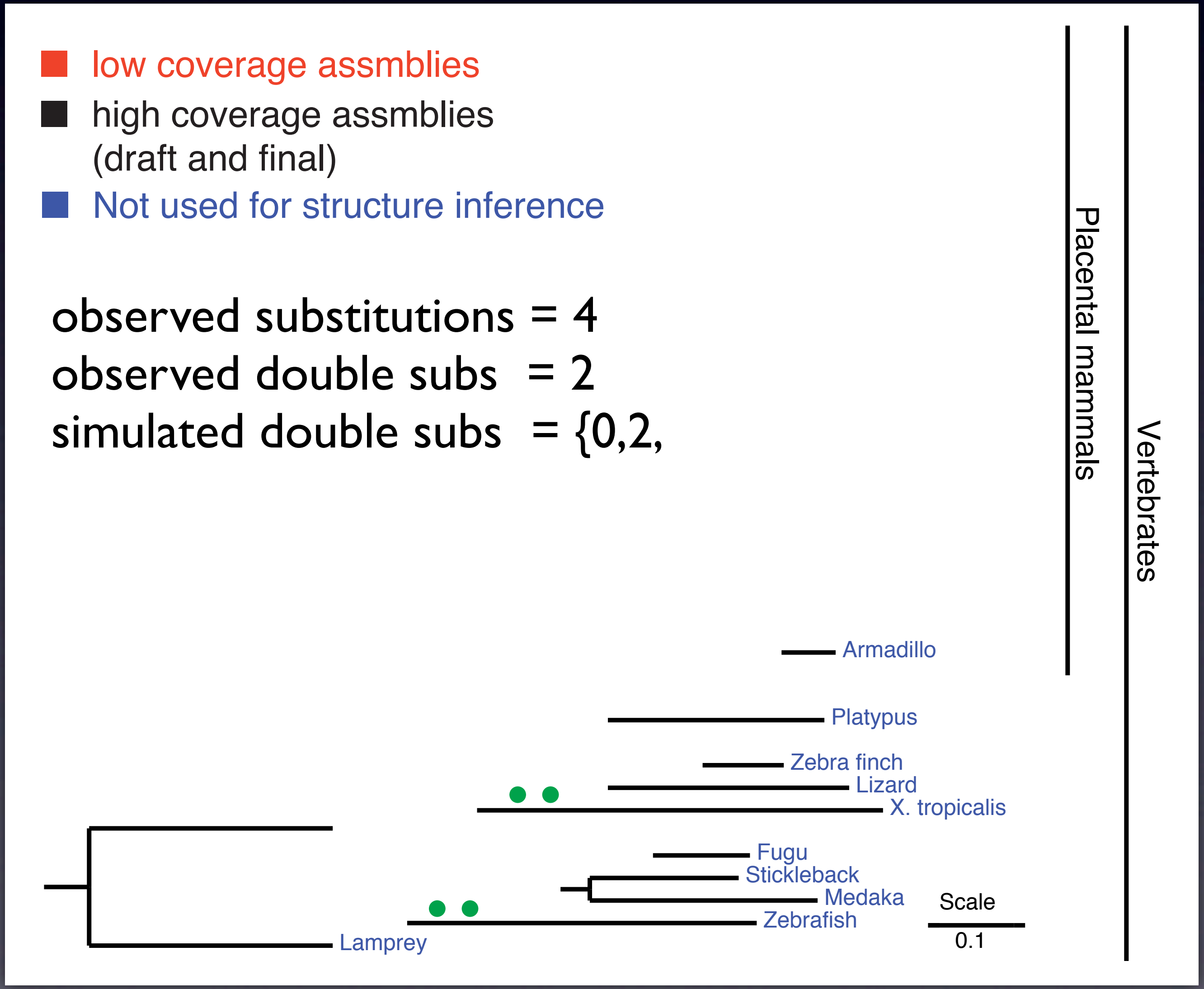
EvoP method

Question: how surprising is the observed number of double substitutions?

- Monte Carlo approach:
  - Simulate iid substitutions across columns on phylogeny.
  - Count double substitutions given structure.
  - Estimate P-value as fraction simulations with at least as many double substitutions.

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

Ten vertebrates genomes not used for structure inference



🟥 low coverage assmblies

⬛ high coverage assmblies (draft and final)

🟦 Not used for structure inference

observed substitutions = 4
observed double subs = 2
simulated double subs = {0,2,

Brian J. Parker, et al. New families of human regulatory RNA structures identified by comparative analysis of vertebrate genomes. Genome Research (2011).

Significance evaluation using additional vertebrate genomes

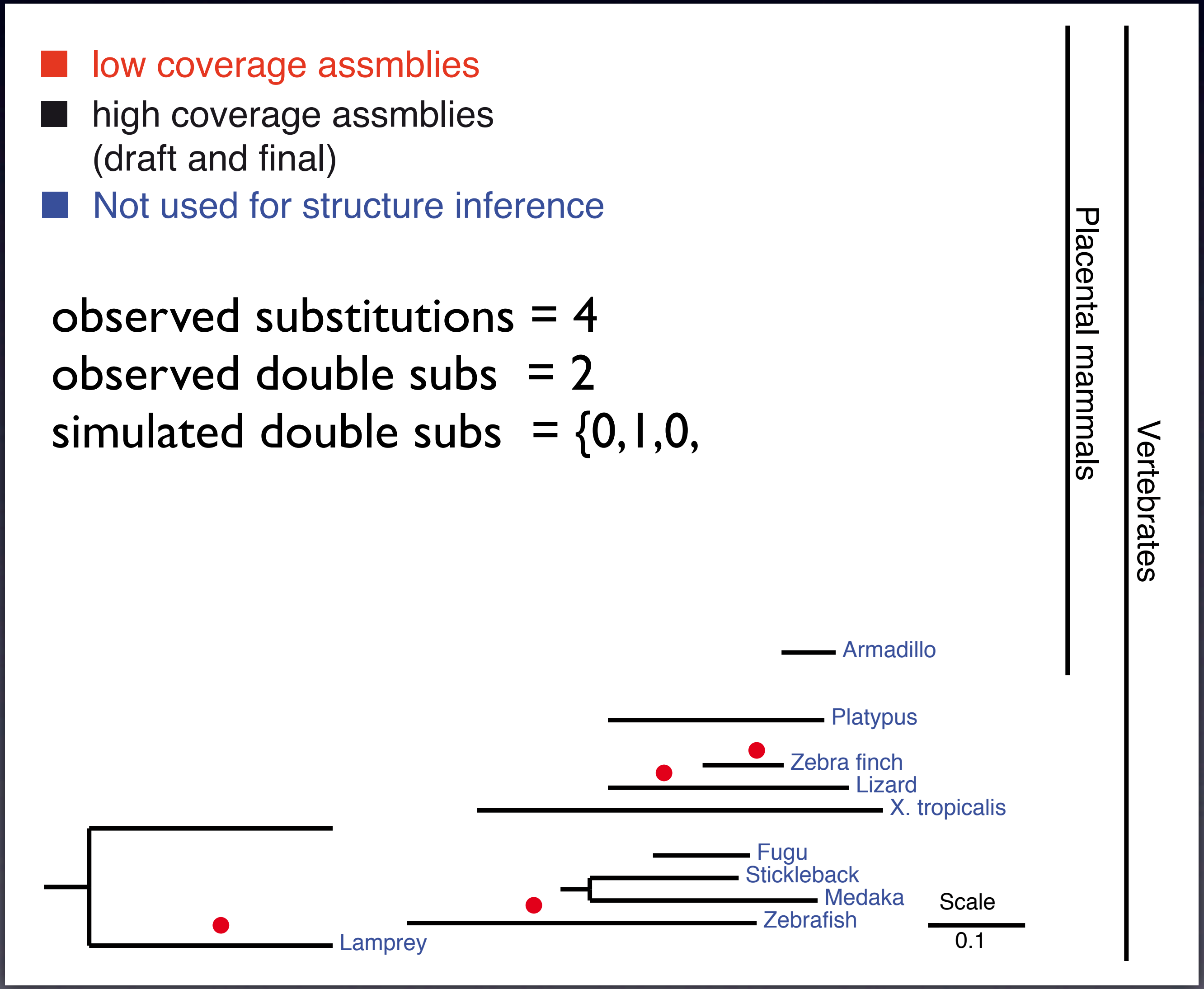
EvoP method

Question: how surprising is the observed number of double substitutions?

- Monte Carlo approach:
  - Simulate iid substitutions across columns on phylogeny.
  - Count double substitutions given structure.
  - Estimate P-value as fraction simulations with at least as many double substitutions.

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

Ten vertebrates genomes not used for structure inference



🟥 low coverage assmblies
⬛ high coverage assmblies (draft and final)
🟦 Not used for structure inference

observed substitutions = 4
observed double subs = 2
simulated double subs = {0,1,0,

Brian J. Parker, et al. New families of human regulatory RNA structures identified by comparative analysis of vertebrate genomes. Genome Research (2011).

Significance evaluation using additional vertebrate genomes

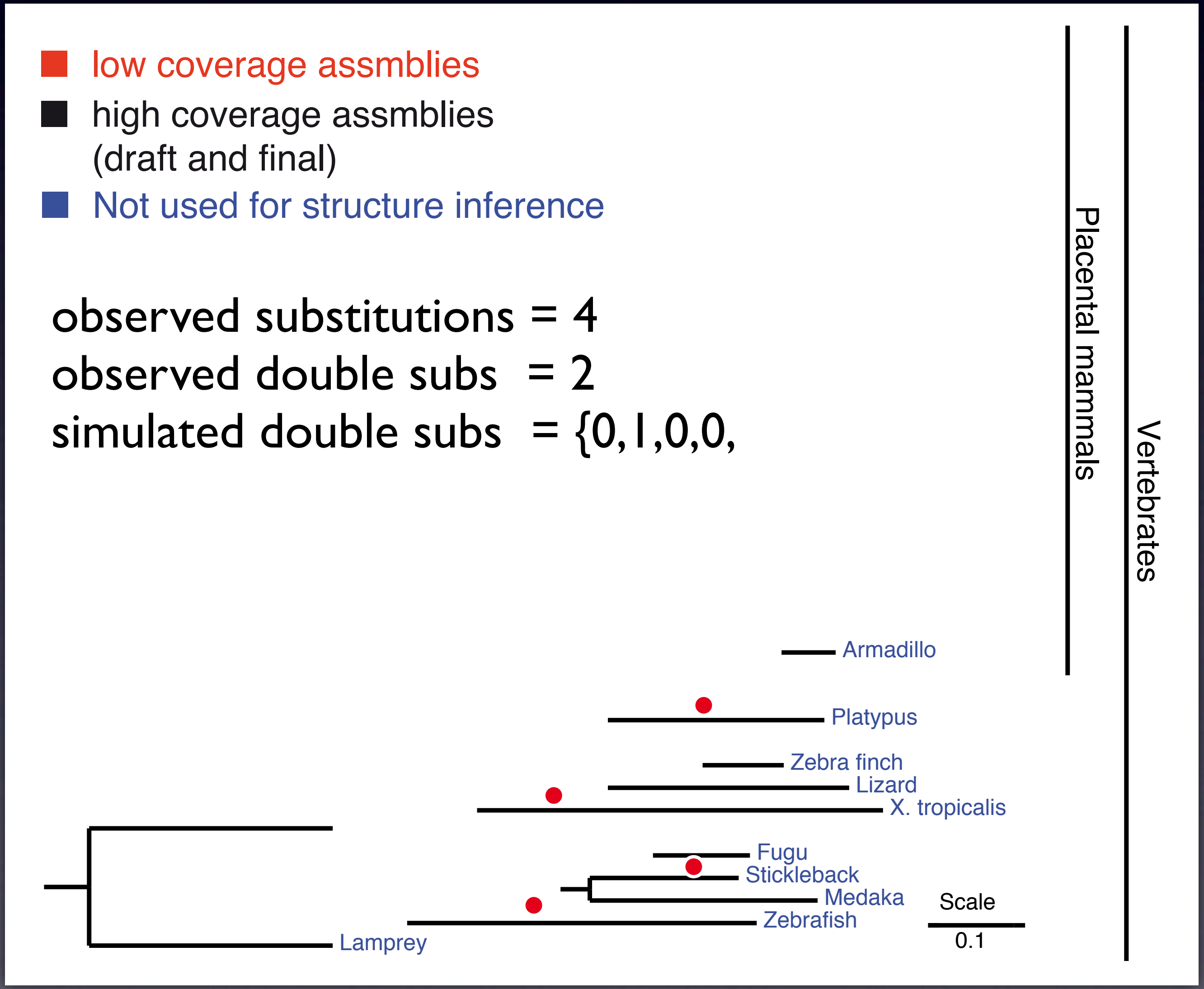
EvoP method

Question: how surprising is the observed number of double substitutions?

- Monte Carlo approach:
  - Simulate iid substitutions across columns on phylogeny.
  - Count double substitutions given structure.
  - Estimate P-value as fraction simulations with at least as many double substitutions.

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

Ten vertebrates genomes not used for structure inference



🟥 low coverage assmblies

⬛ high coverage assmblies (draft and final)

🟦 Not used for structure inference

observed substitutions = 4
observed double subs = 2
simulated double subs = {0,1,0,0,

Brian J. Parker, et al. New families of human regulatory RNA structures identified by comparative analysis of vertebrate genomes. Genome Research (2011).

Significance evaluation using additional vertebrate genomes

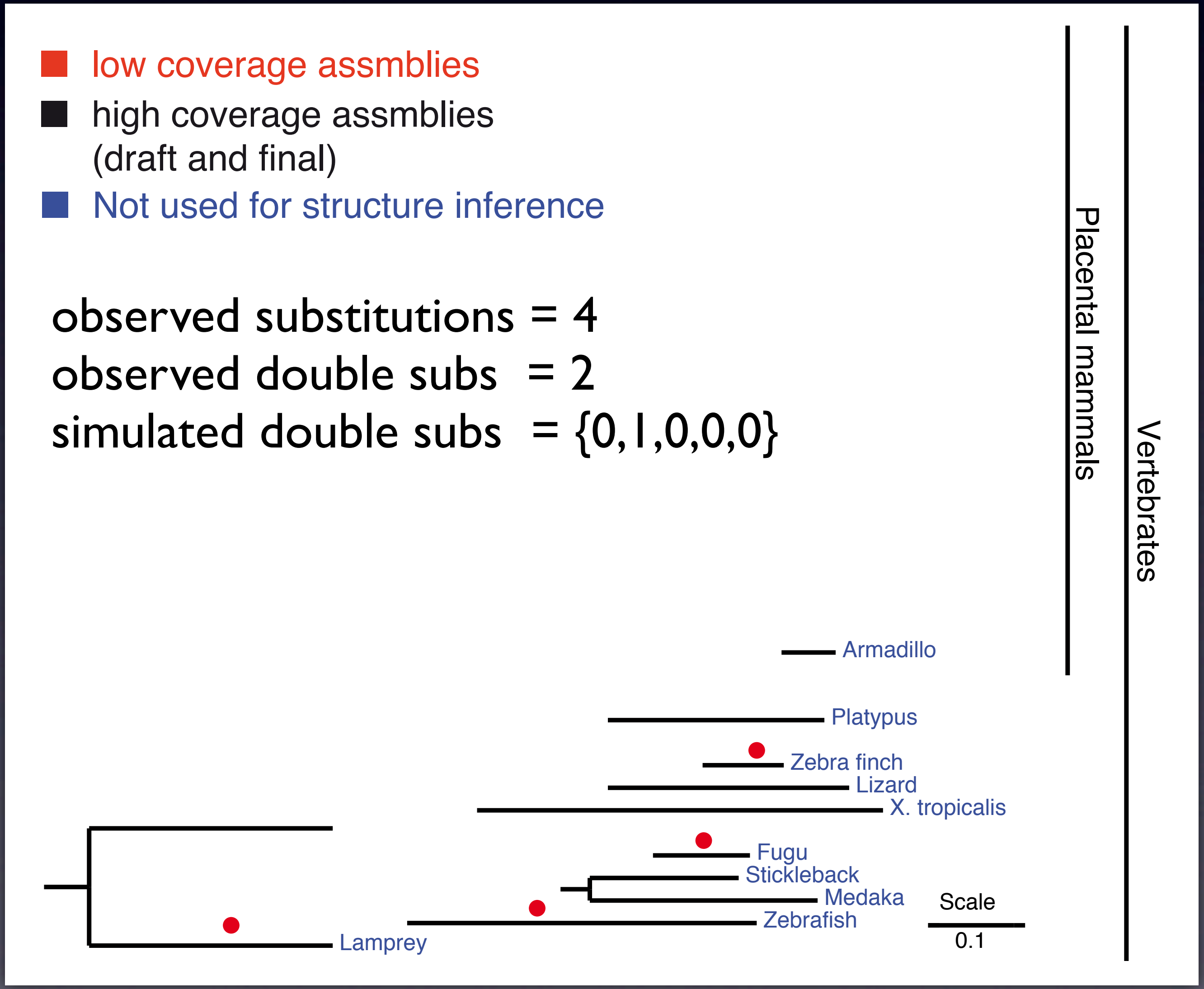
EvoP method

Question: how surprising is the observed number of double substitutions?

- Monte Carlo approach:
  - Simulate iid substitutions across columns on phylogeny.
  - Count double substitutions given structure.
  - Estimate P-value as fraction simulations with at least as many double substitutions.

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

Ten vertebrates genomes not used for structure inference



observed substitutions = 4
observed double subs = 2
simulated double subs = {0,1,0,0,0}

Brian J. Parker, et al. New families of human regulatory RNA structures identified by comparative analysis of vertebrate genomes. Genome Research (2011).

Significance evaluation using additional vertebrate genomes

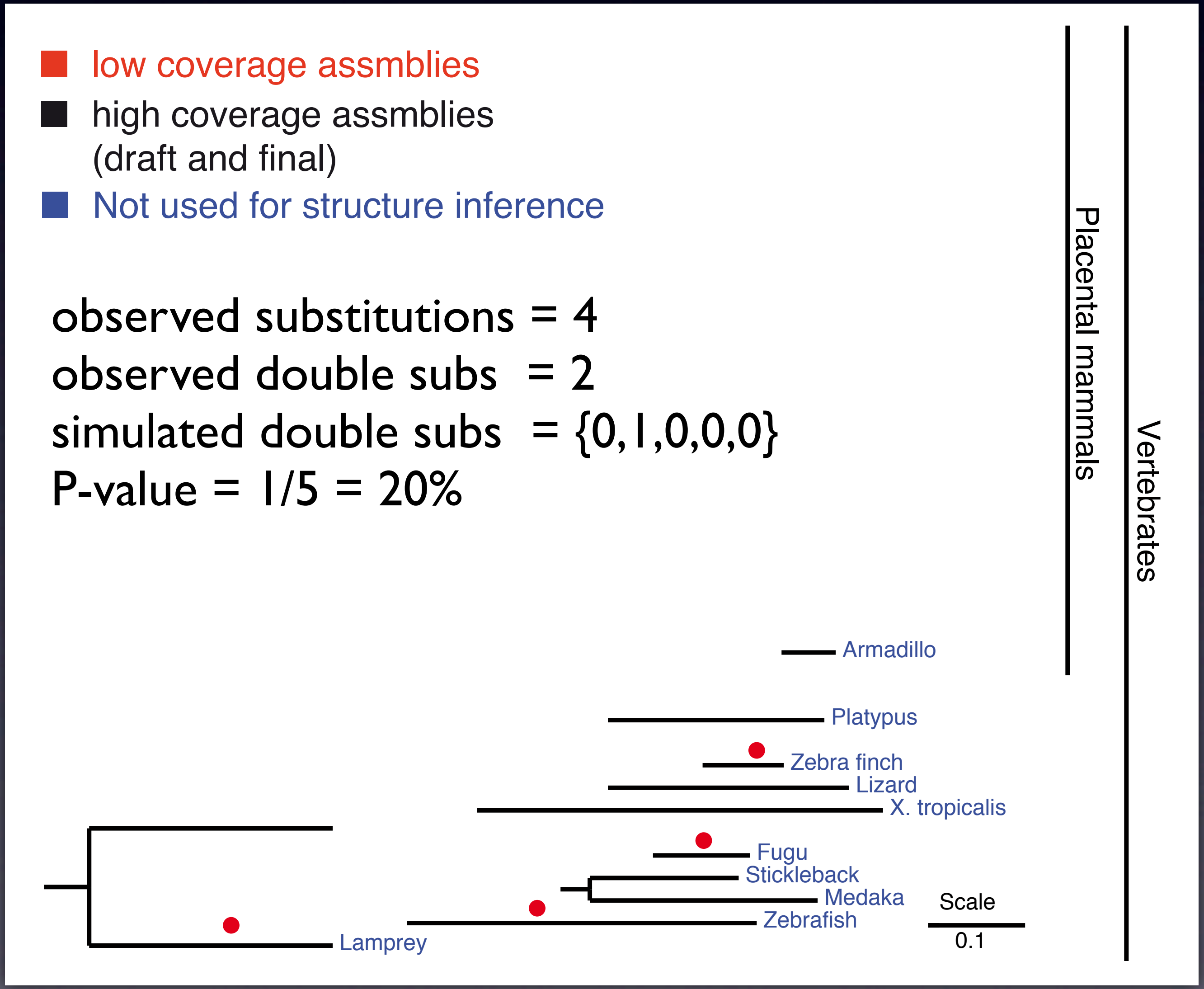
EvoP method

Question: how surprising is the observed number of double substitutions?

- Monte Carlo approach:
  - Simulate iid substitutions across columns on phylogeny.
  - Count double substitutions given structure.
  - Estimate P-value as fraction simulations with at least as many double substitutions.

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

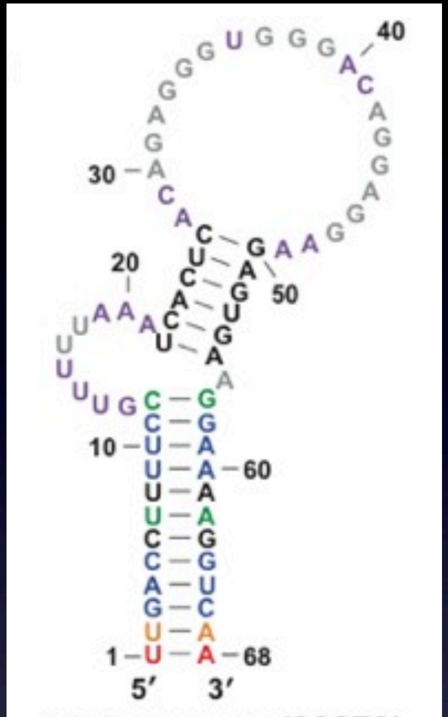
Ten vertebrates genomes not used for structure inference



observed substitutions = 4
observed double subs = 2
simulated double subs = {0,1,0,0,0}
P-value = 1/5 = 20%

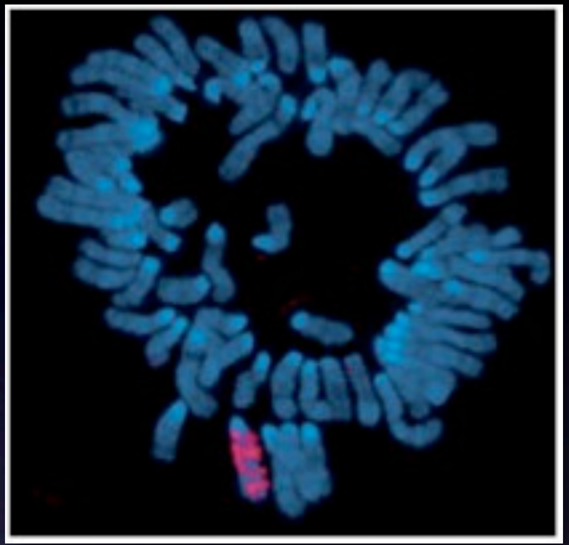
Brian J. Parker, et al. New families of human regulatory RNA structures identified by comparative analysis of vertebrate genomes. Genome Research (2011).

New structure in XIST

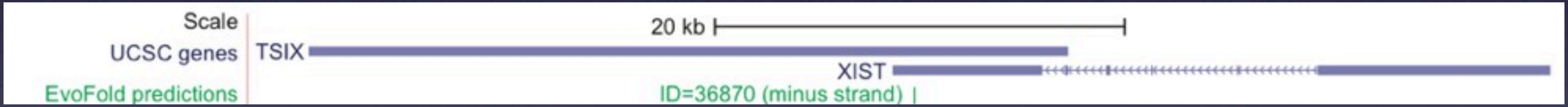


XIST structure (36870)

XIST Chromatin regulation

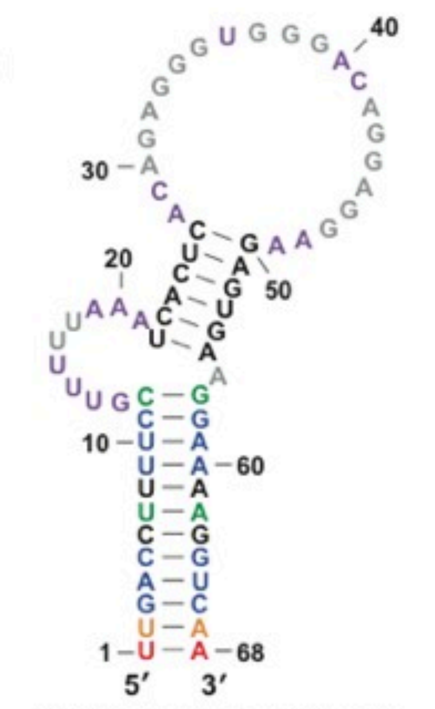


Ng et al. EMBO reports 8, 1, 34–39 (2007).



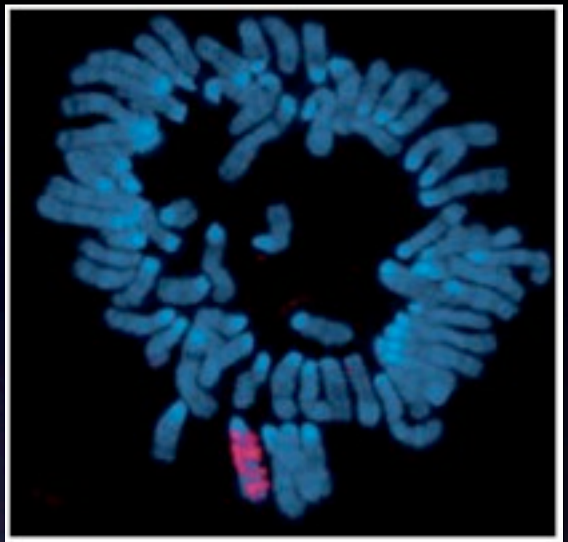
Brian J. Parker et al. New families of human regulatory RNA structures identified by comparative analysis of vertebrate genomes. Genome Research (2011).

New structure in XIST

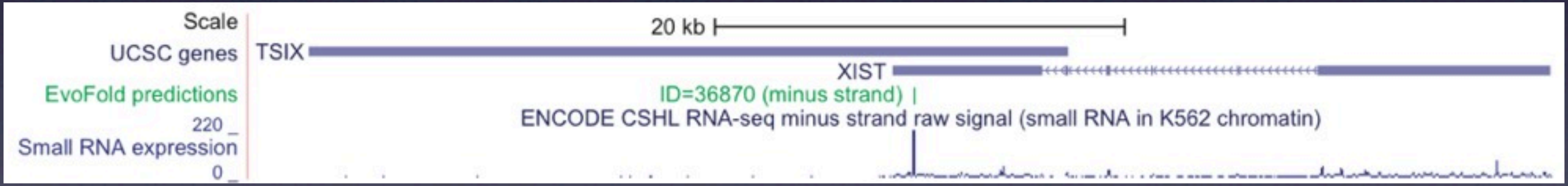


XIST structure (36870)

XIST Chromatin regulation

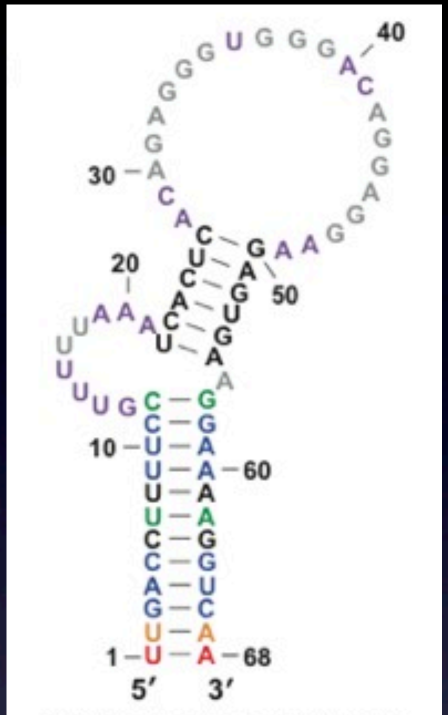


Ng et al. EMBO reports 8, 1, 34–39 (2007).



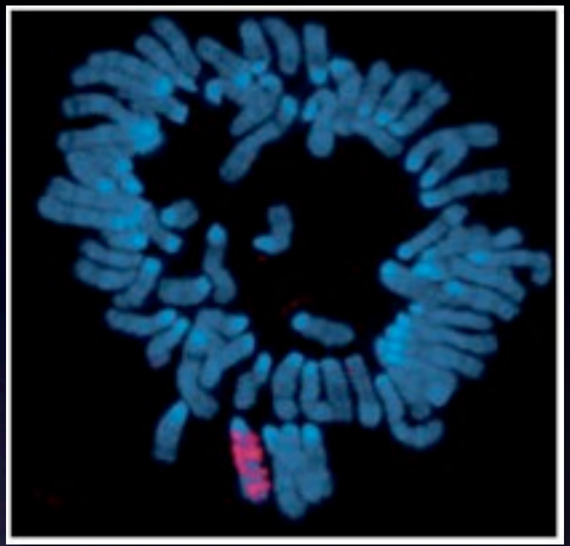
Brian J. Parker et al. New families of human regulatory RNA structures identified by comparative analysis of vertebrate genomes. Genome Research (2011).

New structure in XIST

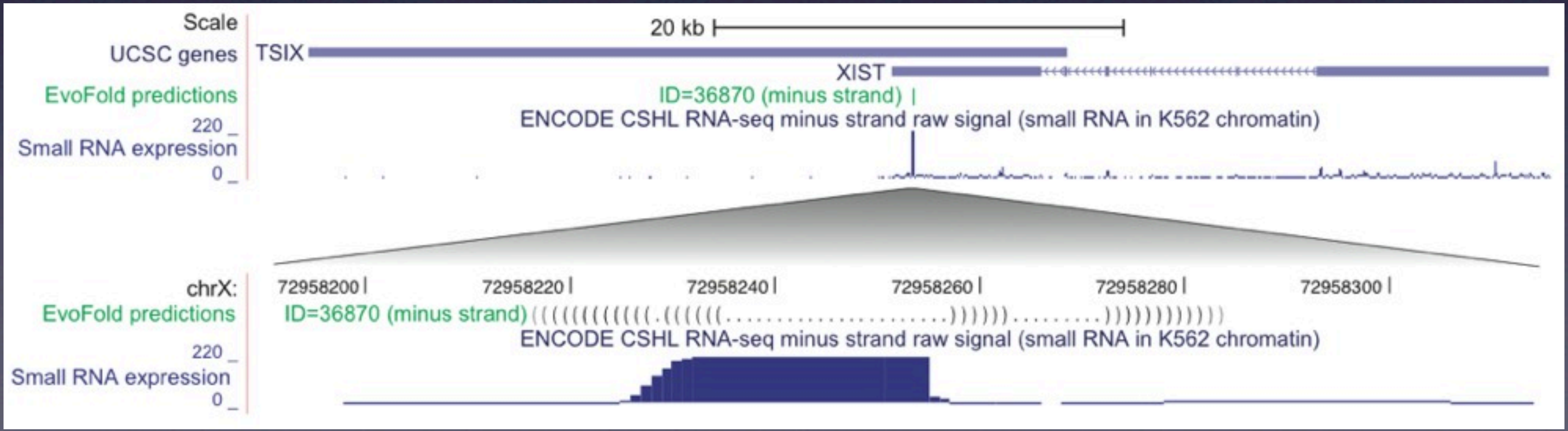


XIST structure (36870)

XIST Chromatin regulation



Ng et al. EMBO reports 8, 1, 34–39 (2007).

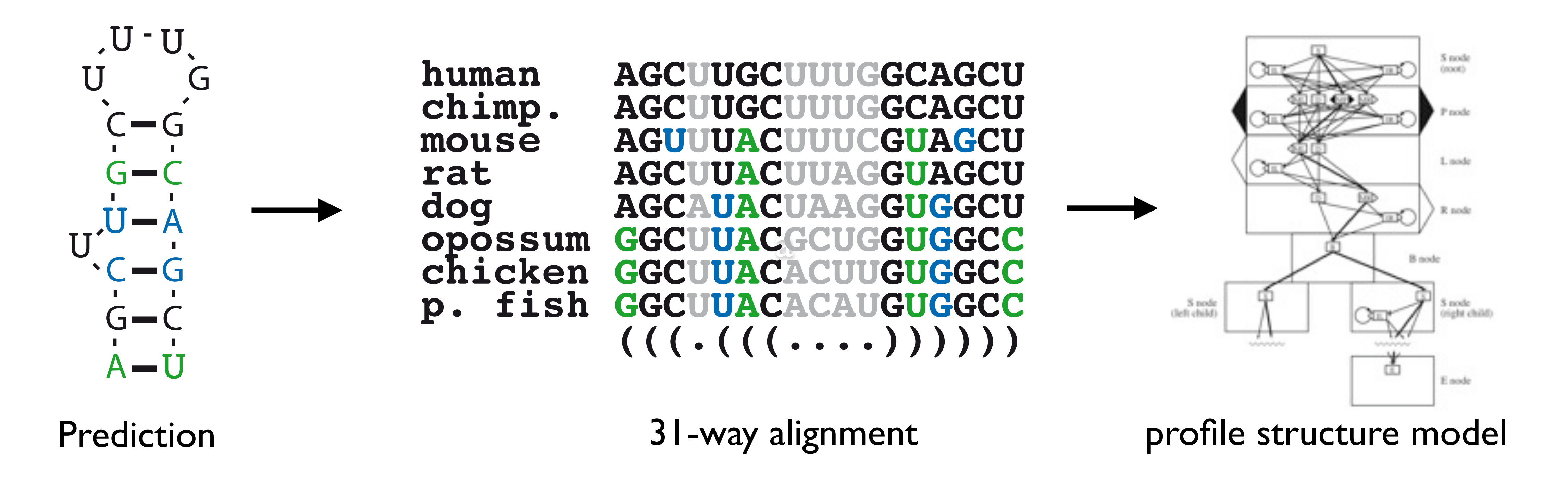

Scale
20 kb
UCSC genes
TSIX
XIST
EvoFold predictions
ID=36870 (minus strand)
ENCODE CSHL RNA-seq minus strand raw signal (small RNA in K562 chromatin)
220
Small RNA expression
0
chrX:
72958200
72958220
72958240
72958260
72958280
72958300
EvoFold predictions
ID=36870 (minus strand) ((((((((((((.((((((......................))))))..........))))))))))))
ENCODE CSHL RNA-seq minus strand raw signal (small RNA in K562 chromatin)
220
Small RNA expression
0

Brian J. Parker et al. New families of human regulatory RNA structures identified by comparative analysis of vertebrate genomes. Genome Research (2011).

Family classification

Similarity measure

Build profile -SCFGs / co-variance models for every prediction



```
human    AGCUUGCUUUGGCAGCU
chimp.   AGCUUGCUUUGGCAGCU
mouse    AGUUUACUUUCGUAGCU
rat      AGCUUACUUAGGUAGCU
dog      AGCAUACUAAGGUGGCU
opossum  GGCUUACGCUGGUGGCC
chicken  GGCUUACACUUGUGGCC
p. fish  GGCUUACACAUGUGGCC
         (((.(((....))))))
```

Prediction → 31-way alignment → profile structure model

Models made with Infernal tools from Sean Eddy's group.

Structure model figure: Durbin et al, Biological Sequence Analysis, Cambridge University Press.
Nawrocki EP, Kolbe DL, Eddy SR. 2009. Infernal 1.0: inference of RNA alignments. Bioinformatics 25: 1335–1337.

(Dis)similarity measure

Initially, we wanted to use Kullback-Liebler divergence (KL):

$$D_{KL}\left(M_1 \parallel M_2\right) = \sum_i P_{M_1}(i) \log \frac{P_{M_1}(i)}{P_{M_2}(i)}$$

(Dis)similarity measure

Initially, we wanted to use Kullback-Liebler divergence (KL):

$$D_{KL}\left(M_1 \parallel M_2\right) = \sum_i P_{M_1}(i) \log \frac{P_{M_1}(i)}{P_{M_2}(i)}$$

We couldn’t compute and resorted to sampling:

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

(Dis)similarity measure

Initially, we wanted to use Kullback-Liebler divergence (KL):

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

We couldn’t compute and resorted to sampling:

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

Still slow. Approximated by one sample only - human sequence from training alignment. Also replaced probabilities with (Infernal) normalized scores:

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

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

(Dis)similarity measure

Initially, we wanted to use Kullback-Liebler divergence (KL):

$$D_{KL}\left(M_1 \parallel M_2\right) = \sum_i P_{M_1}(i) \log \frac{P_{M_1}(i)}{P_{M_2}(i)}$$

We couldn’t compute and resorted to sampling:

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

Still slow. Approximated by one sample only - human sequence from training alignment. Also replaced probabilities with (Infernal) normalized scores:

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

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

Problem: Models of different complexities have different false positive rates. Hence replace score with E-score.

$$\tilde{D}(M_1 \| M_2) = E\left(S\left(seq_1^{human}, M_2\right)\right) - E\left(S\left(seq_1^{human}, M_1\right)\right)$$

(Dis)similarity measure

Initially, we wanted to use Kullback-Liebler divergence (KL):

$$D_{KL}(M_1 \| M_2) = \sum_i P_{M_1}(i) \log \frac{P_{M_1}(i)}{P_{M_2}(i)}$$

We couldn’t compute and resorted to sampling:

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

Still slow. Approximated by one sample only - human sequence from training alignment.
Also replaced probabilities with (Infernal) normalized scores:

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

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

Problem: Models of different complexities have different false positive rates.
Hence replace score with E-score.

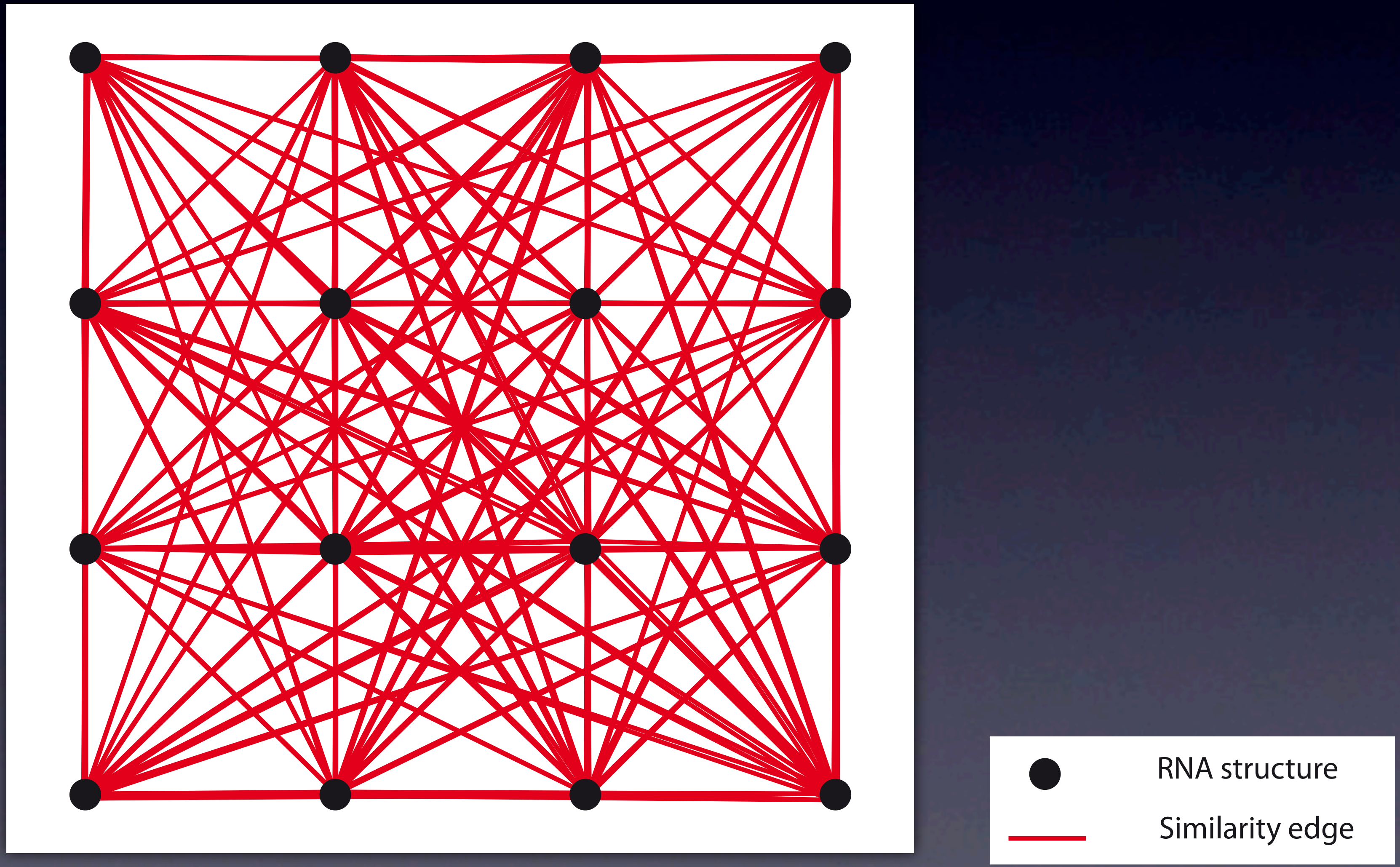
$$\tilde{D}(M_1 \| M_2) = E(S(seq_1^{human}, M_2)) - E(S(seq_1^{human}, M_1))$$

Finally, be conservative and symmetrize by max:

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

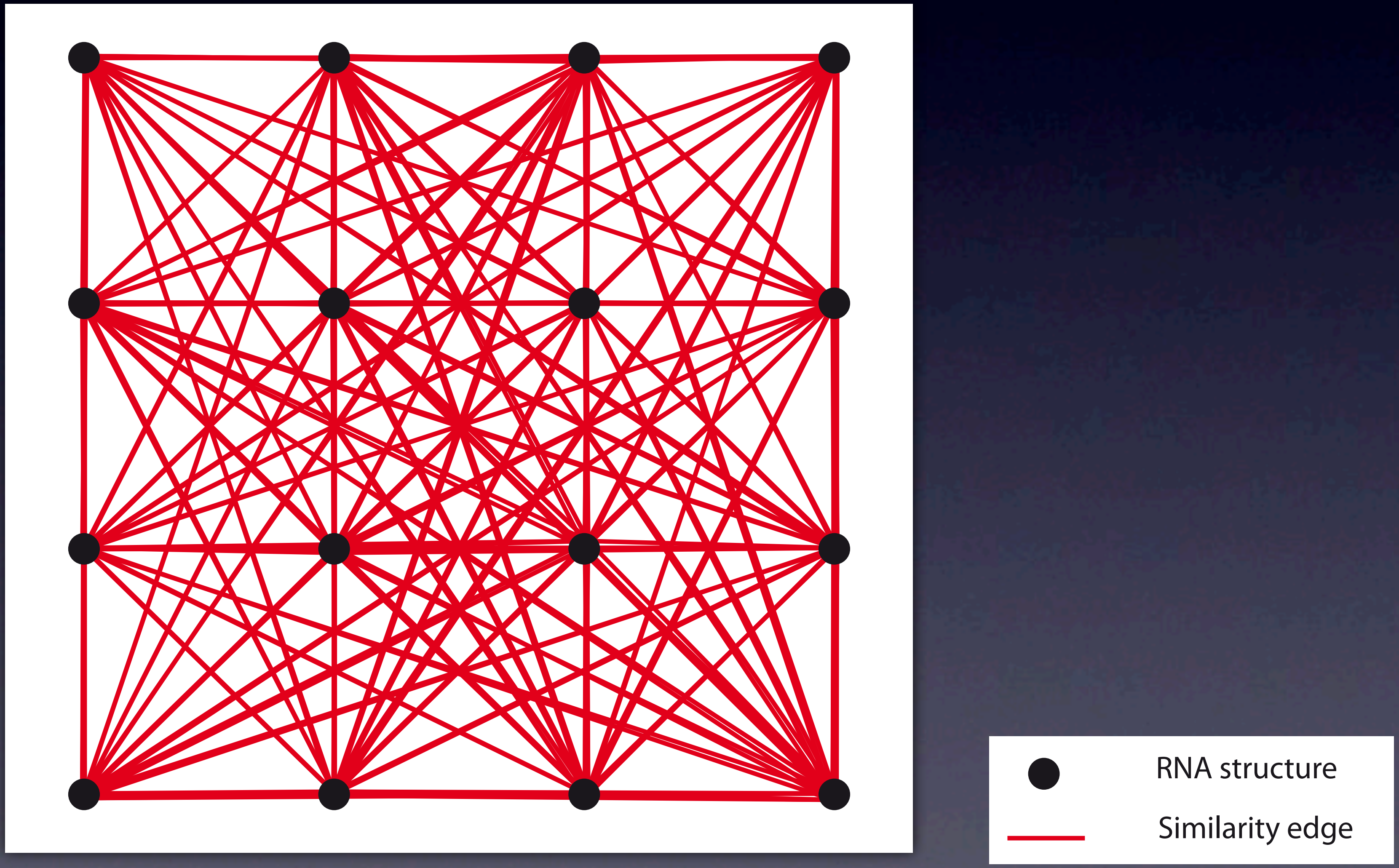
Graph-based clustering into families

All pairwise comparisons



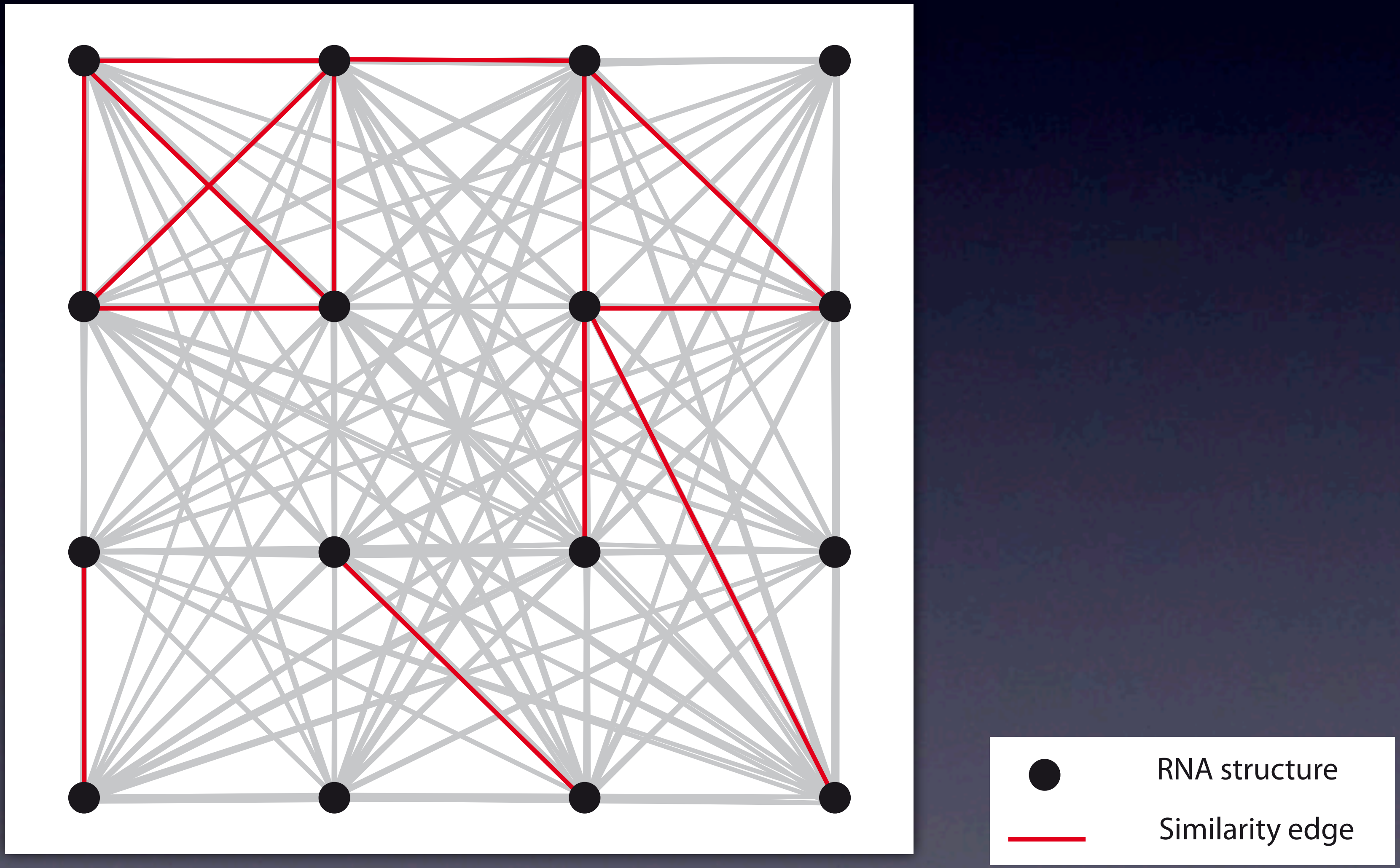
Graph-based clustering into families

Threshold on similarity significance



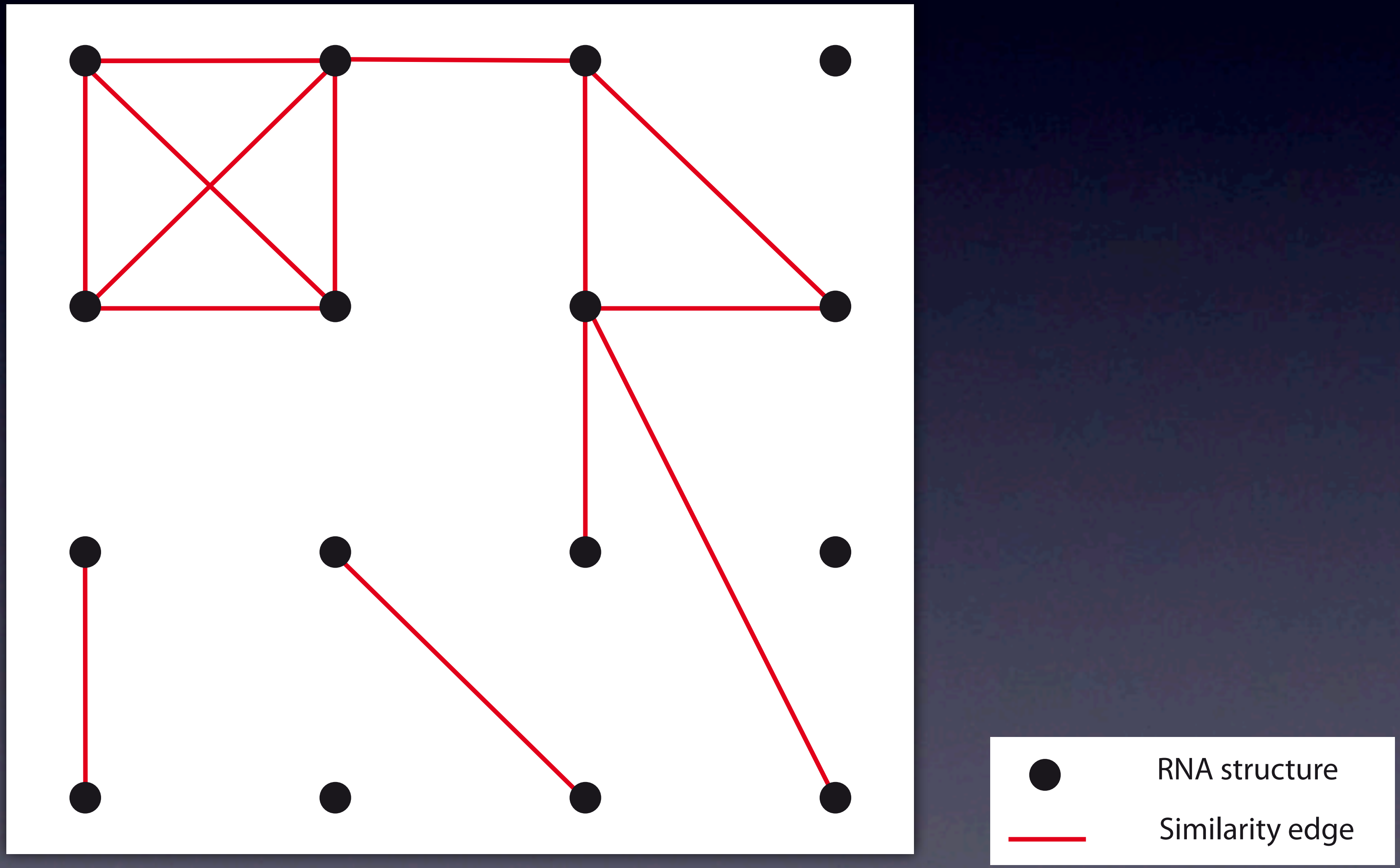
Graph-based clustering into families

Threshold on similarity significance



Graph-based clustering into families

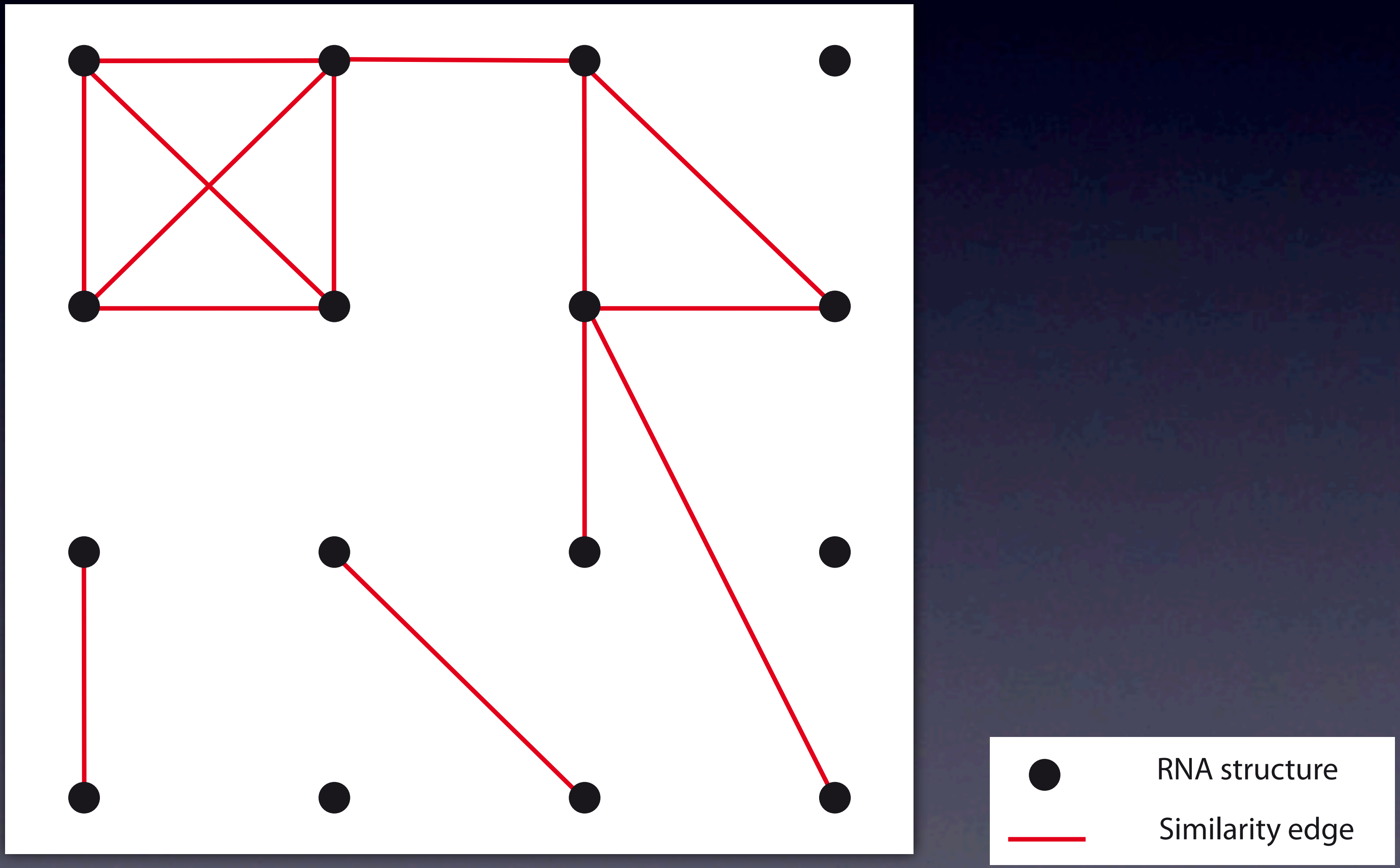
Threshold on similarity significance



Hartuv and Shamir (2000)

Graph-based clustering into families

Identify highly connected subgraph (HCS)



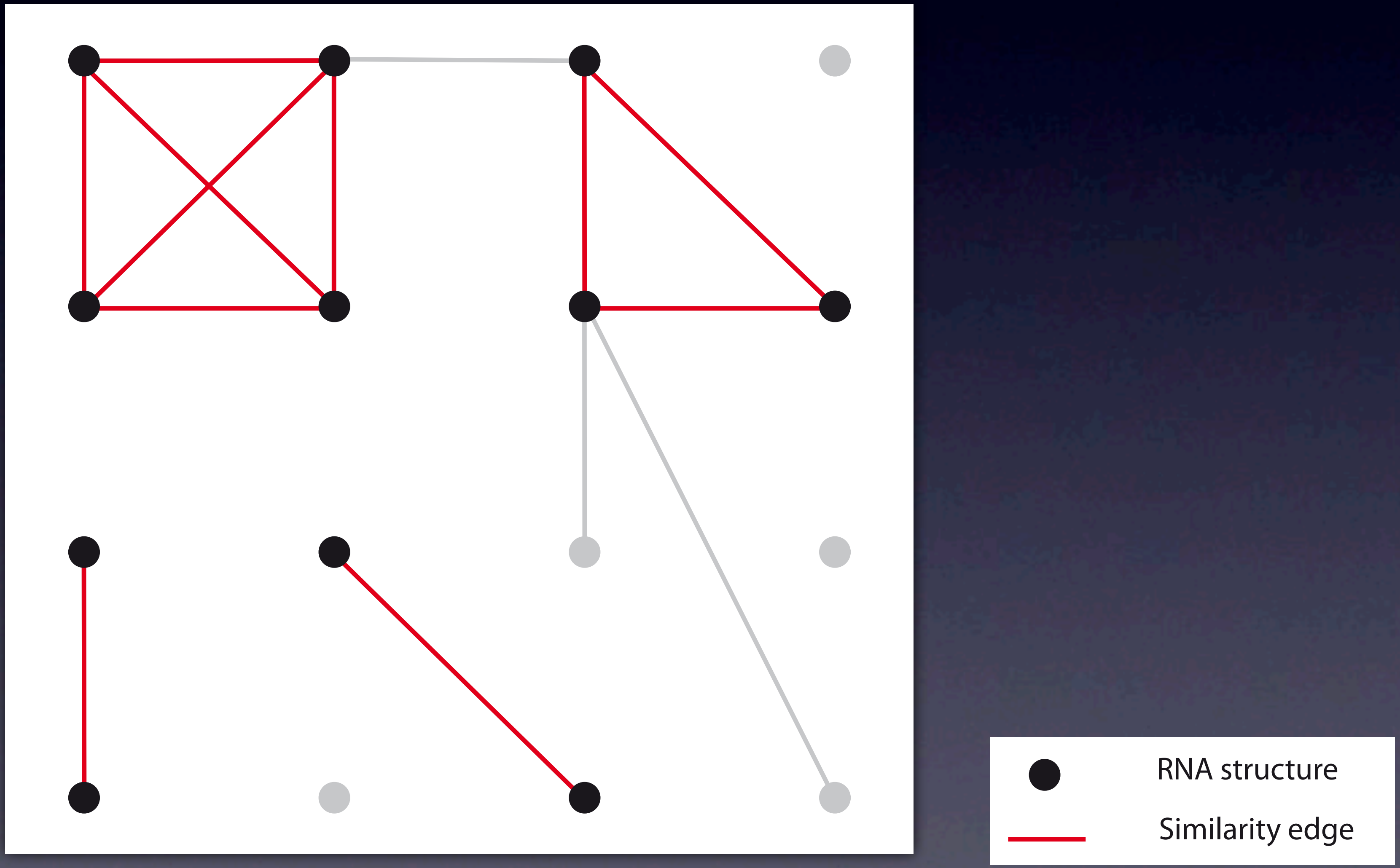
Highly connected subgraph:

- edge connectivity < half number of vertices
- subgraph size two

Hartuv and Shamir (2000)

Graph-based clustering into families

Identify highly connected subgraph (HCS)



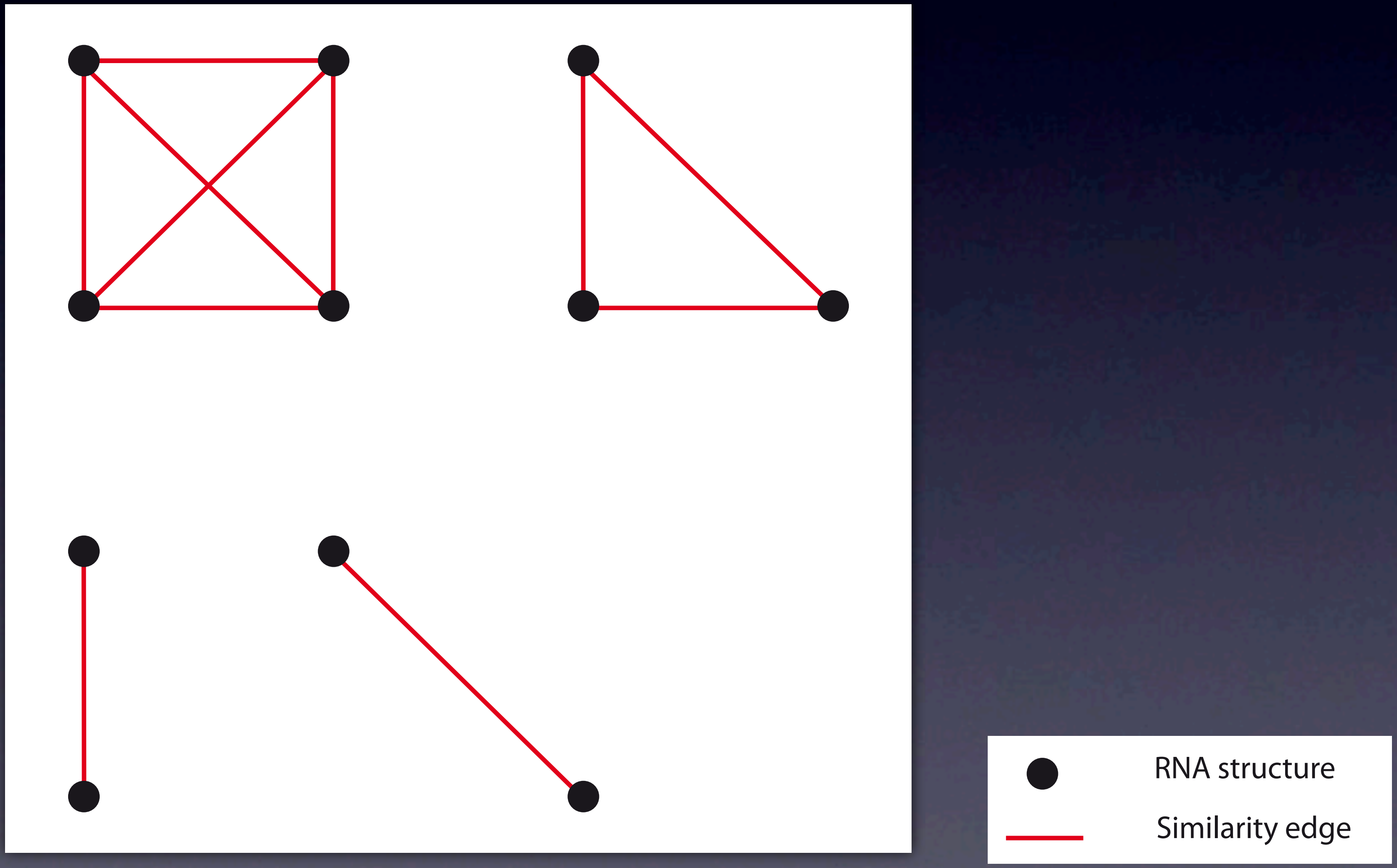
Highly connected subgraph:

- edge connectivity < half number of vertices
- subgraph size two

Hartuv and Shamir (2000)

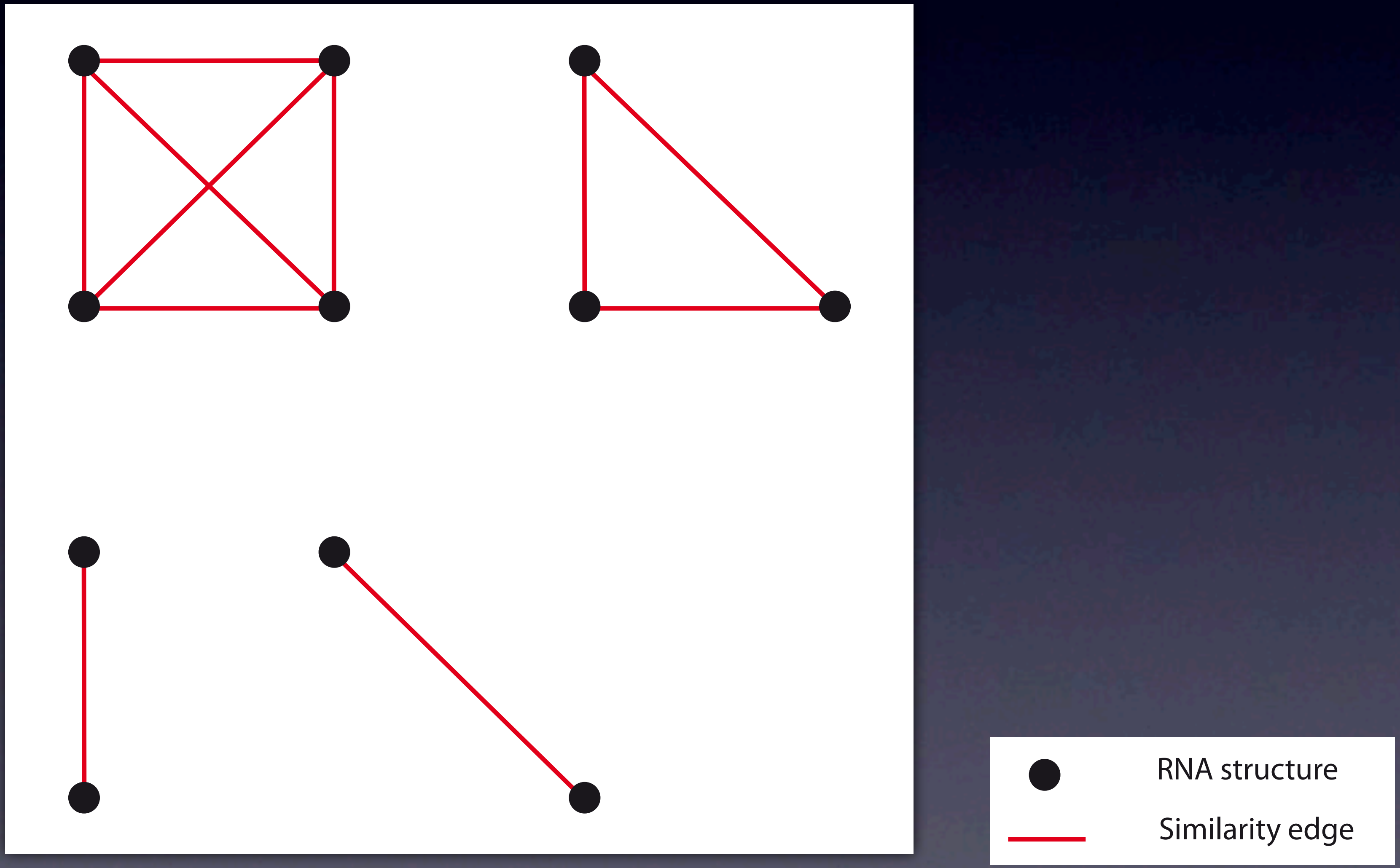
Graph-based clustering into families

Identify highly connected subsets (HCS)

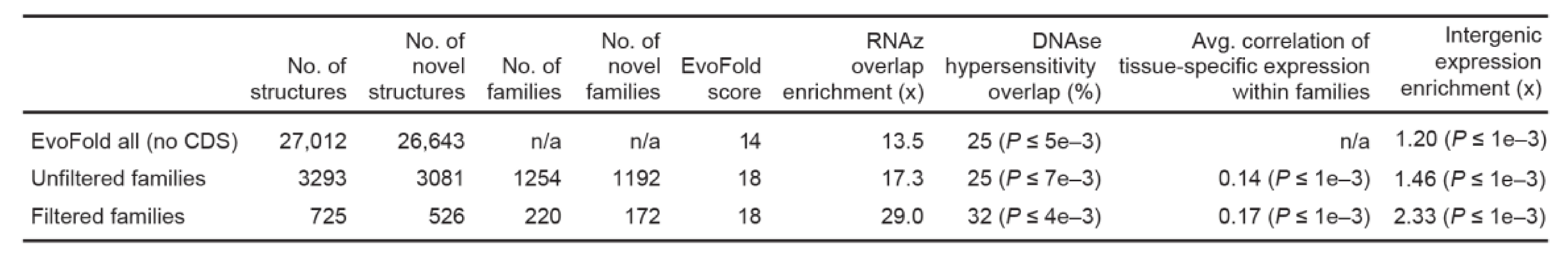


Graph-based clustering into families

Final family candidates



Family prediction overview

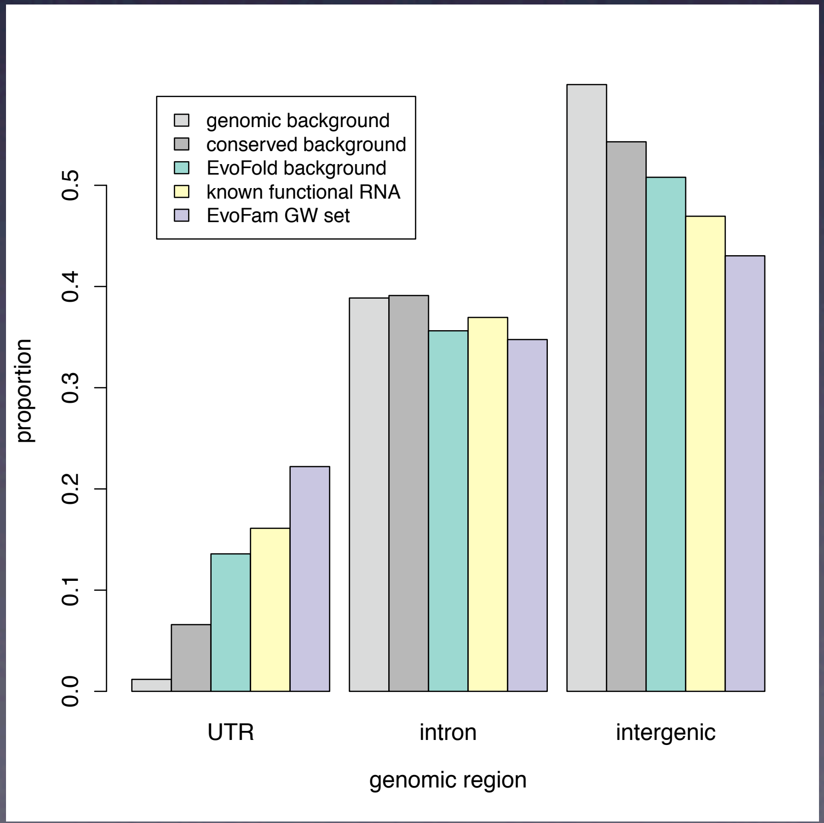


	No. of structures	No. of novel structures	No. of families	No. of novel families	EvoFold score	RNAz overlap enrichment (x)	DNAse hypersensitivity overlap (%)	Avg. correlation of tissue-specific expression within families	Intergenic expression enrichment (x)
EvoFold all (no CDS)	27,012	26,643	n/a	n/a	14	13.5	25 ($P \leq$ 5e–3)	n/a	1.20 ($P \leq$ 1e–3)
Unfiltered families	3293	3081	1254	1192	18	17.3	25 ($P \leq$ 7e–3)	0.14 ($P \leq$ 1e–3)	1.46 ($P \leq$ 1e–3)
Filtered families	725	526	220	172	18	29.0	32 ($P \leq$ 4e–3)	0.17 ($P \leq$ 1e–3)	2.33 ($P \leq$ 1e–3)

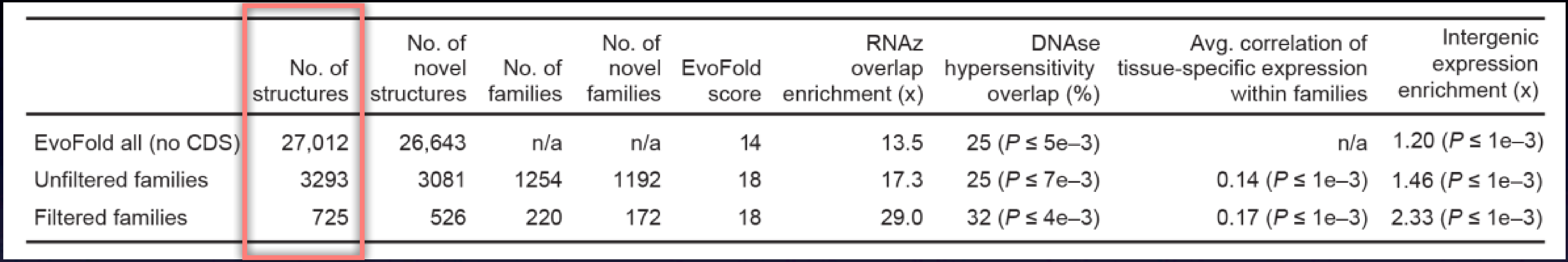
Filtered families have either:

- EvoP test < 0.05
- Region enrichment test < 0.005
- GO enrichment test relative to EvoFold background < 0.01
- Mean structure length > 11 base pairs

Genomic distribution



Family prediction overview

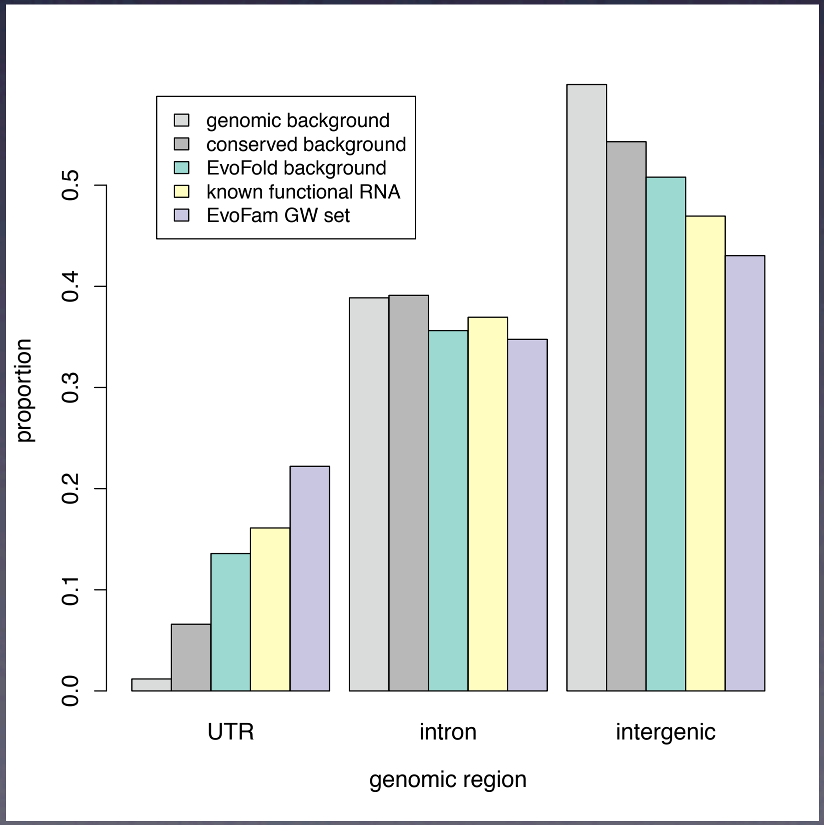


	No. of structures	No. of novel structures	No. of families	No. of novel families	EvoFold score	RNAz overlap enrichment (x)	DNAse hypersensitivity overlap (%)	Avg. correlation of tissue-specific expression within families	Intergenic expression enrichment (x)
EvoFold all (no CDS)	27,012	26,643	n/a	n/a	14	13.5	25 ($P \leq$ 5e–3)	n/a	1.20 ($P \leq$ 1e–3)
Unfiltered families	3293	3081	1254	1192	18	17.3	25 ($P \leq$ 7e–3)	0.14 ($P \leq$ 1e–3)	1.46 ($P \leq$ 1e–3)
Filtered families	725	526	220	172	18	29.0	32 ($P \leq$ 4e–3)	0.17 ($P \leq$ 1e–3)	2.33 ($P \leq$ 1e–3)

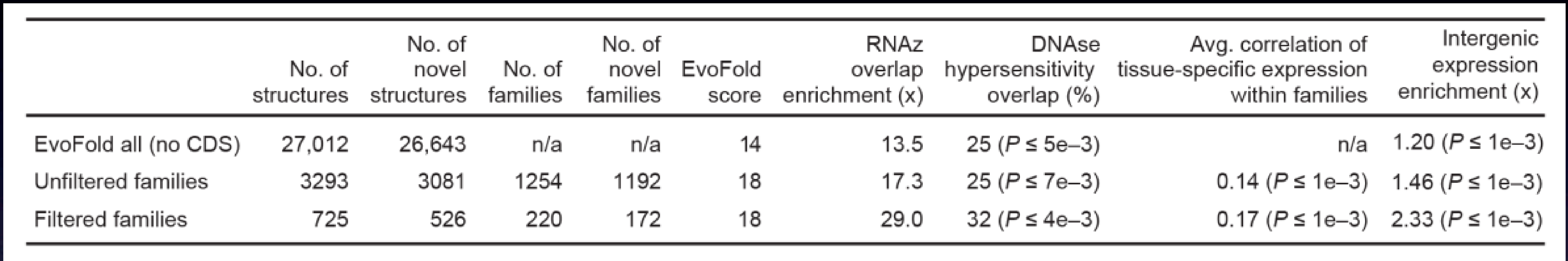
Filtered families have either:

- EvoP test < 0.05
- Region enrichment test < 0.005
- GO enrichment test relative to EvoFold background < 0.01
- Mean structure length > 11 base pairs

Genomic distribution



Family prediction overview

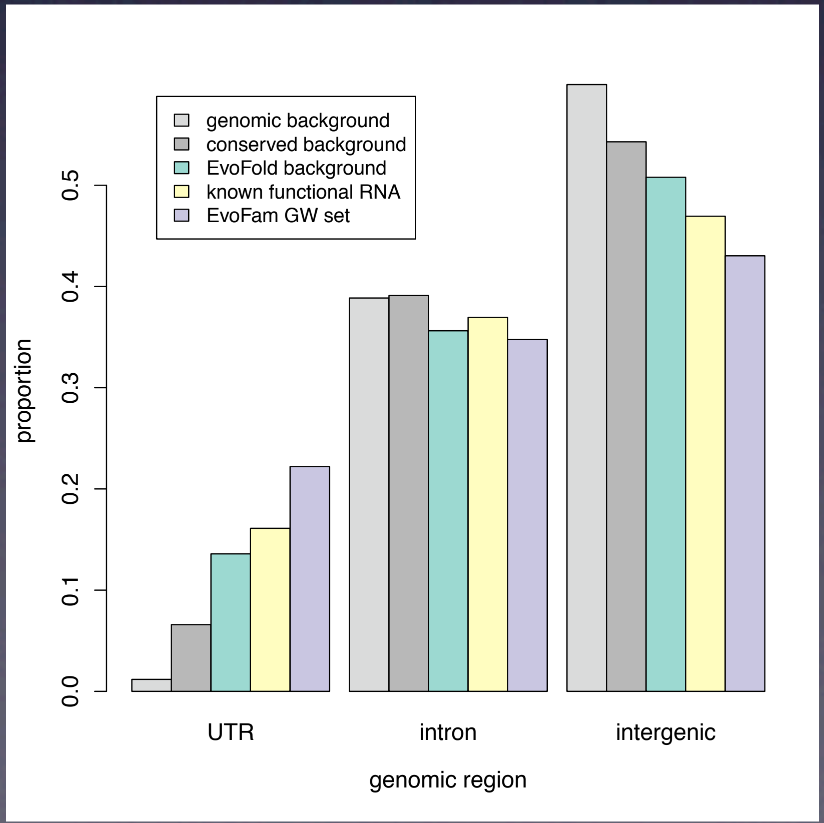


	No. of structures	No. of novel structures	No. of families	No. of novel families	EvoFold score	RNAz overlap enrichment (x)	DNAse hypersensitivity overlap (%)	Avg. correlation of tissue-specific expression within families	Intergenic expression enrichment (x)
EvoFold all (no CDS)	27,012	26,643	n/a	n/a	14	13.5	25 ($P \leq$ 5e–3)	n/a	1.20 ($P \leq$ 1e–3)
Unfiltered families	3293	3081	1254	1192	18	17.3	25 ($P \leq$ 7e–3)	0.14 ($P \leq$ 1e–3)	1.46 ($P \leq$ 1e–3)
Filtered families	725	526	220	172	18	29.0	32 ($P \leq$ 4e–3)	0.17 ($P \leq$ 1e–3)	2.33 ($P \leq$ 1e–3)

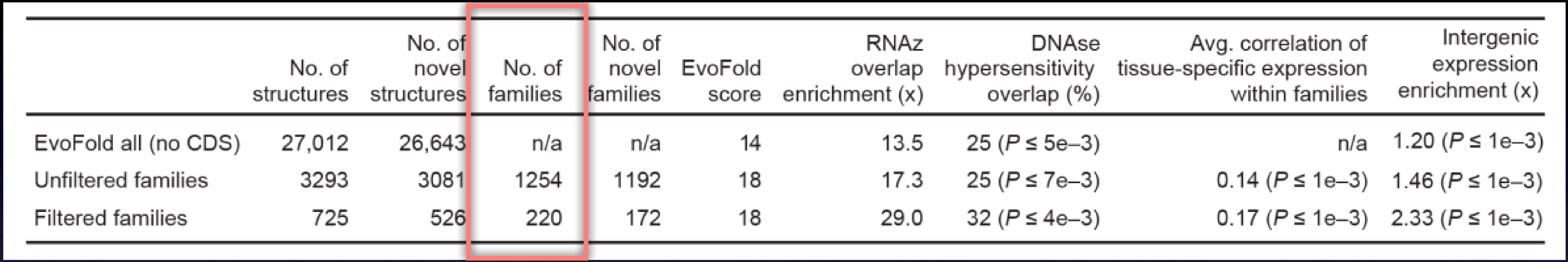
Filtered families have either:

- EvoP test < 0.05
- Region enrichment test < 0.005
- GO enrichment test relative to EvoFold background < 0.01
- Mean structure length > 11 base pairs

Genomic distribution



Family prediction overview

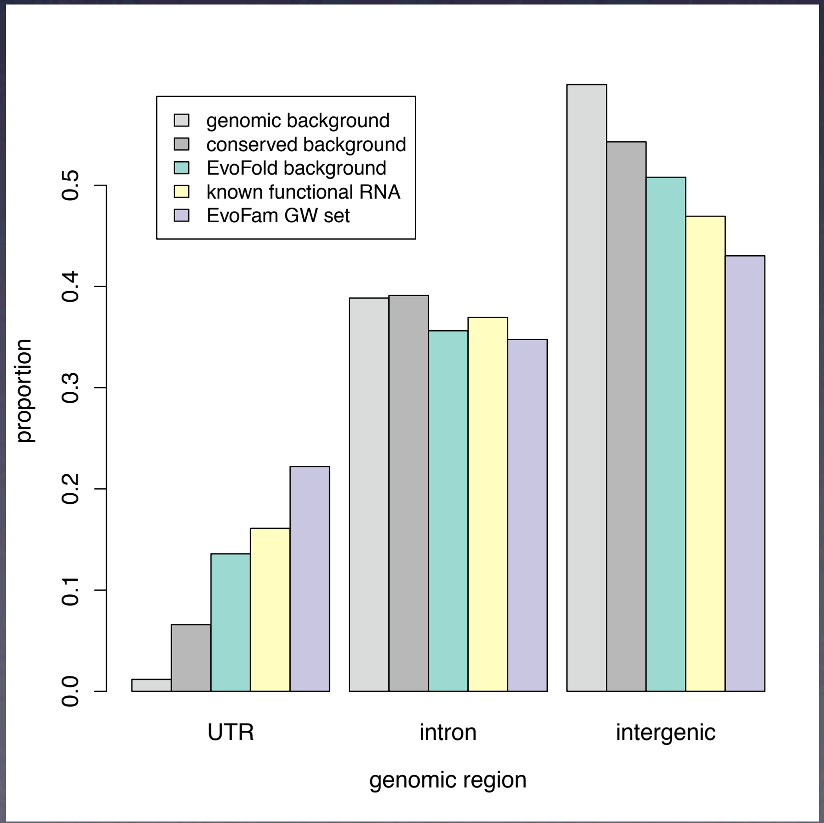


	No. of structures	No. of novel structures	No. of families	No. of novel families	EvoFold score	RNAz overlap enrichment (x)	DNAse hypersensitivity overlap (%)	Avg. correlation of tissue-specific expression within families	Intergenic expression enrichment (x)
EvoFold all (no CDS)	27,012	26,643	n/a	n/a	14	13.5	25 ($P \le$ 5e–3)	n/a	1.20 ($P \le$ 1e–3)
Unfiltered families	3293	3081	1254	1192	18	17.3	25 ($P \le$ 7e–3)	0.14 ($P \le$ 1e–3)	1.46 ($P \le$ 1e–3)
Filtered families	725	526	220	172	18	29.0	32 ($P \le$ 4e–3)	0.17 ($P \le$ 1e–3)	2.33 ($P \le$ 1e–3)

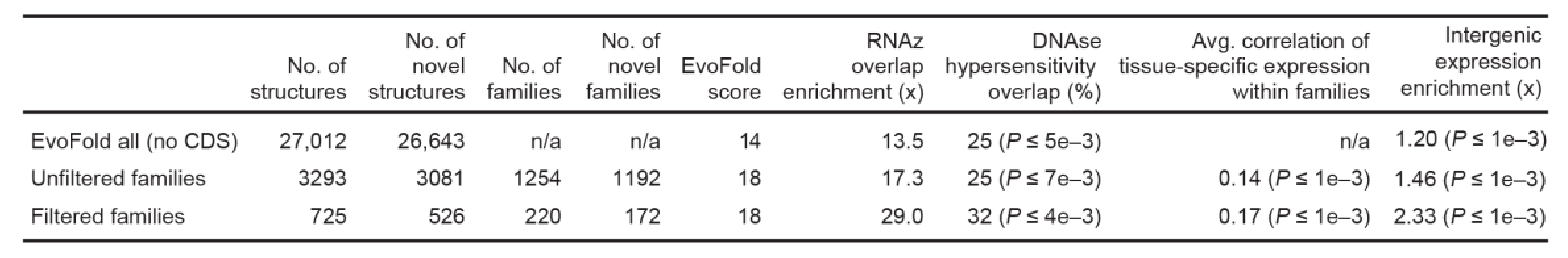
Filtered families have either:

- EvoP test < 0.05
- Region enrichment test < 0.005
- GO enrichment test relative to EvoFold background < 0.01
- Mean structure length > 11 base pairs

Genomic distribution



Family prediction overview

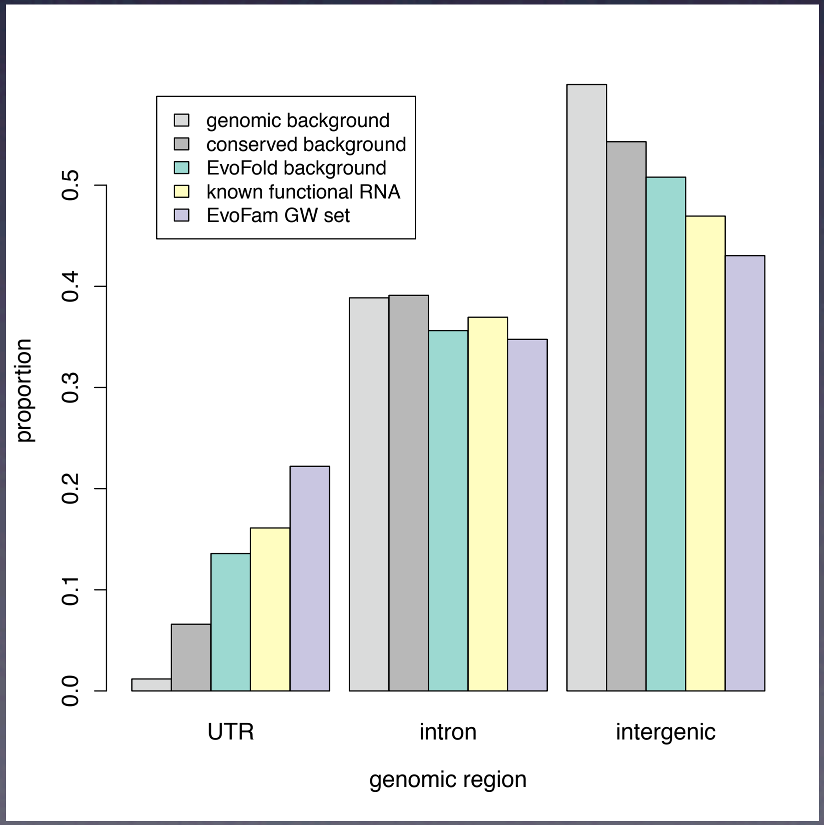


	No. of structures	No. of novel structures	No. of families	No. of novel families	EvoFold score	RNAz overlap enrichment (x)	DNAse hypersensitivity overlap (%)	Avg. correlation of tissue-specific expression within families	Intergenic expression enrichment (x)
EvoFold all (no CDS)	27,012	26,643	n/a	n/a	14	13.5	25 (*P* ≤ 5e–3)	n/a	1.20 (*P* ≤ 1e–3)
Unfiltered families	3293	3081	1254	1192	18	17.3	25 (*P* ≤ 7e–3)	0.14 (*P* ≤ 1e–3)	1.46 (*P* ≤ 1e–3)
Filtered families	725	526	220	172	18	29.0	32 (*P* ≤ 4e–3)	0.17 (*P* ≤ 1e–3)	2.33 (*P* ≤ 1e–3)

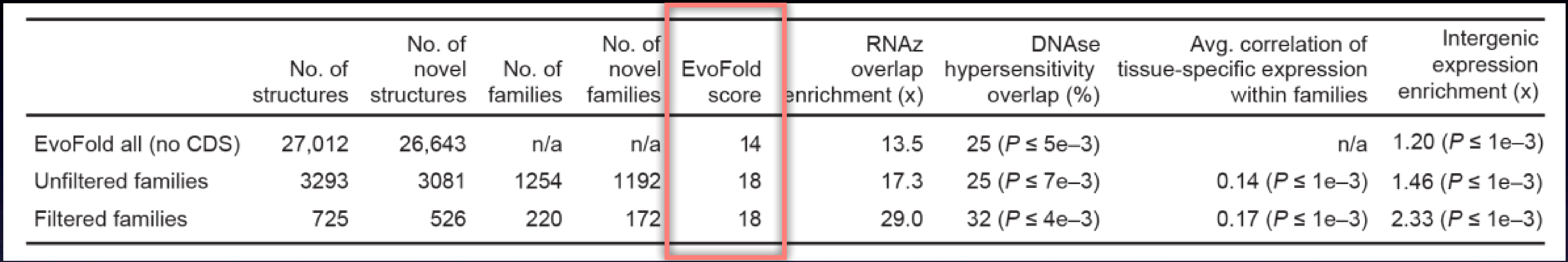
Filtered families have either:

- EvoP test < 0.05
- Region enrichment test < 0.005
- GO enrichment test relative to EvoFold background < 0.01
- Mean structure length > 11 base pairs

Genomic distribution



Family prediction overview

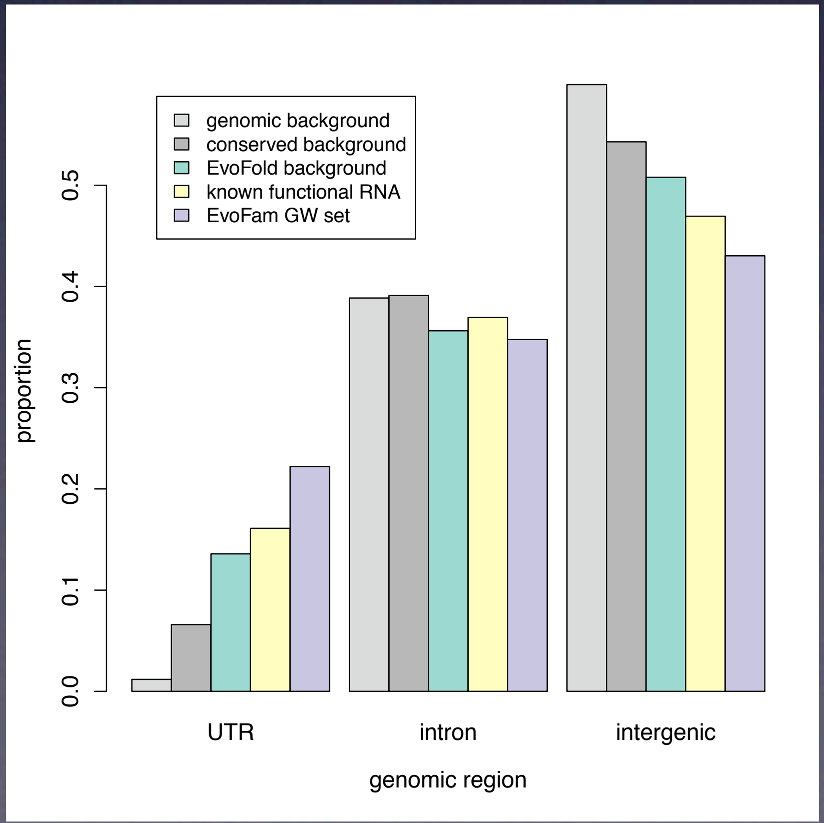


	No. of structures	No. of novel structures	No. of families	No. of novel families	EvoFold score	RNAz overlap enrichment (x)	DNAse hypersensitivity overlap (%)	Avg. correlation of tissue-specific expression within families	Intergenic expression enrichment (x)
EvoFold all (no CDS)	27,012	26,643	n/a	n/a	14	13.5	25 ($P \leq$ 5e–3)	n/a	1.20 ($P \leq$ 1e–3)
Unfiltered families	3293	3081	1254	1192	18	17.3	25 ($P \leq$ 7e–3)	0.14 ($P \leq$ 1e–3)	1.46 ($P \leq$ 1e–3)
Filtered families	725	526	220	172	18	29.0	32 ($P \leq$ 4e–3)	0.17 ($P \leq$ 1e–3)	2.33 ($P \leq$ 1e–3)

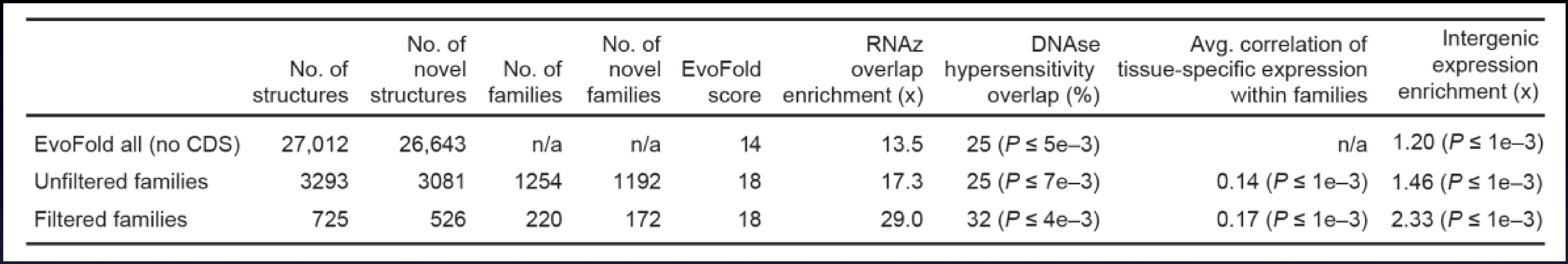
Filtered families have either:

- EvoP test < 0.05
- Region enrichment test < 0.005
- GO enrichment test relative to EvoFold background < 0.01
- Mean structure length > 11 base pairs

Genomic distribution



Family prediction overview

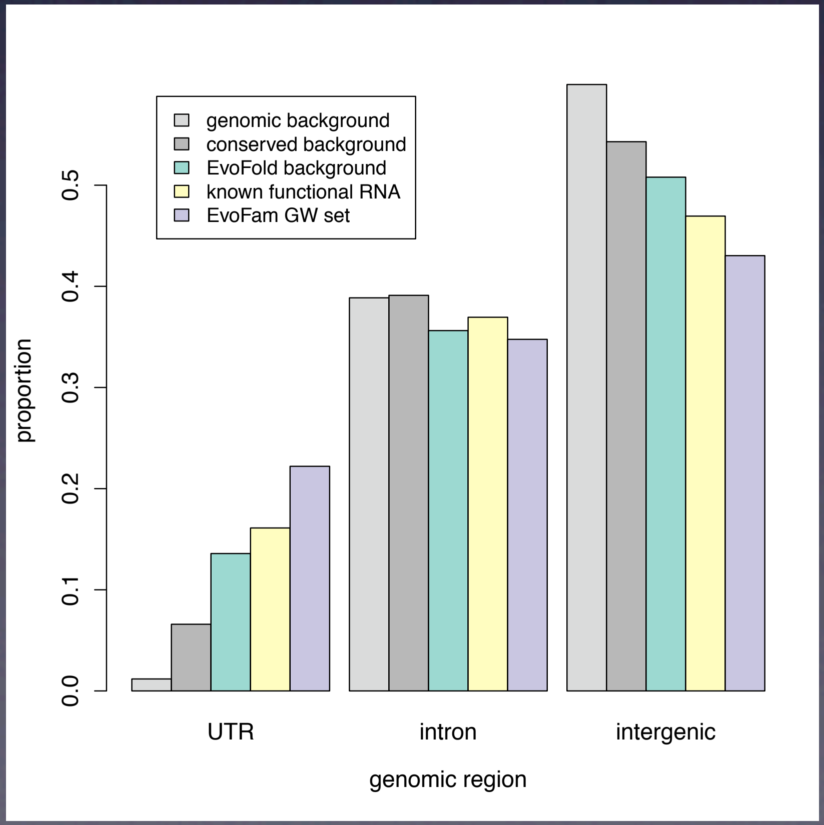


	No. of structures	No. of novel structures	No. of families	No. of novel families	EvoFold score	RNAz overlap enrichment (x)	DNAse hypersensitivity overlap (%)	Avg. correlation of tissue-specific expression within families	Intergenic expression enrichment (x)
EvoFold all (no CDS)	27,012	26,643	n/a	n/a	14	13.5	25 ($P \le$ 5e–3)	n/a	1.20 ($P \le$ 1e–3)
Unfiltered families	3293	3081	1254	1192	18	17.3	25 ($P \le$ 7e–3)	0.14 ($P \le$ 1e–3)	1.46 ($P \le$ 1e–3)
Filtered families	725	526	220	172	18	29.0	32 ($P \le$ 4e–3)	0.17 ($P \le$ 1e–3)	2.33 ($P \le$ 1e–3)

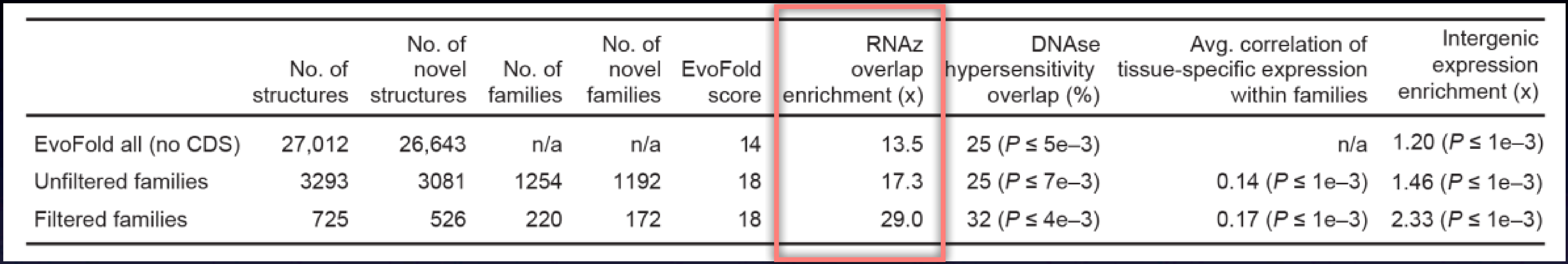
Filtered families have either:

- EvoP test < 0.05
- Region enrichment test < 0.005
- GO enrichment test relative to EvoFold background < 0.01
- Mean structure length > 11 base pairs

Genomic distribution



Family prediction overview

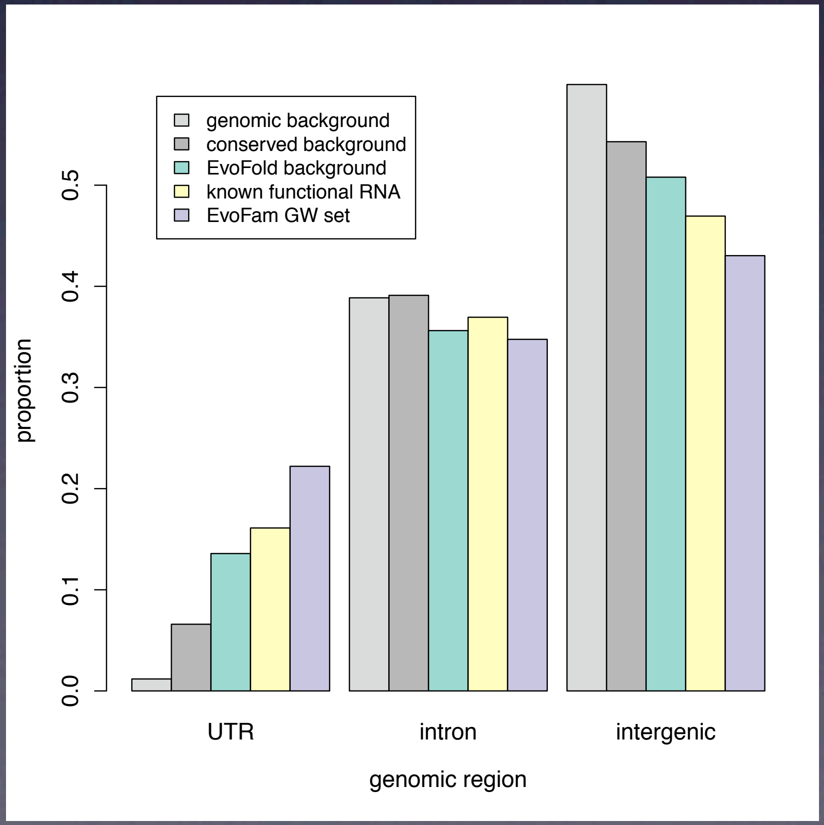


	No. of structures	No. of novel structures	No. of families	No. of novel families	EvoFold score	RNAz overlap enrichment (x)	DNAse hypersensitivity overlap (%)	Avg. correlation of tissue-specific expression within families	Intergenic expression enrichment (x)
EvoFold all (no CDS)	27,012	26,643	n/a	n/a	14	13.5	25 ($P \leq$ 5e–3)	n/a	1.20 ($P \leq$ 1e–3)
Unfiltered families	3293	3081	1254	1192	18	17.3	25 ($P \leq$ 7e–3)	0.14 ($P \leq$ 1e–3)	1.46 ($P \leq$ 1e–3)
Filtered families	725	526	220	172	18	29.0	32 ($P \leq$ 4e–3)	0.17 ($P \leq$ 1e–3)	2.33 ($P \leq$ 1e–3)

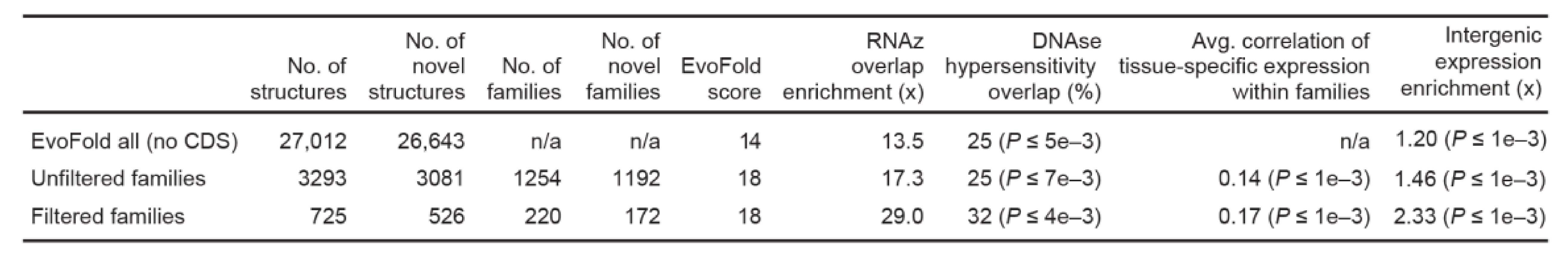
Filtered families have either:

- EvoP test < 0.05
- Region enrichment test < 0.005
- GO enrichment test relative to EvoFold background < 0.01
- Mean structure length > 11 base pairs

Genomic distribution



Family prediction overview

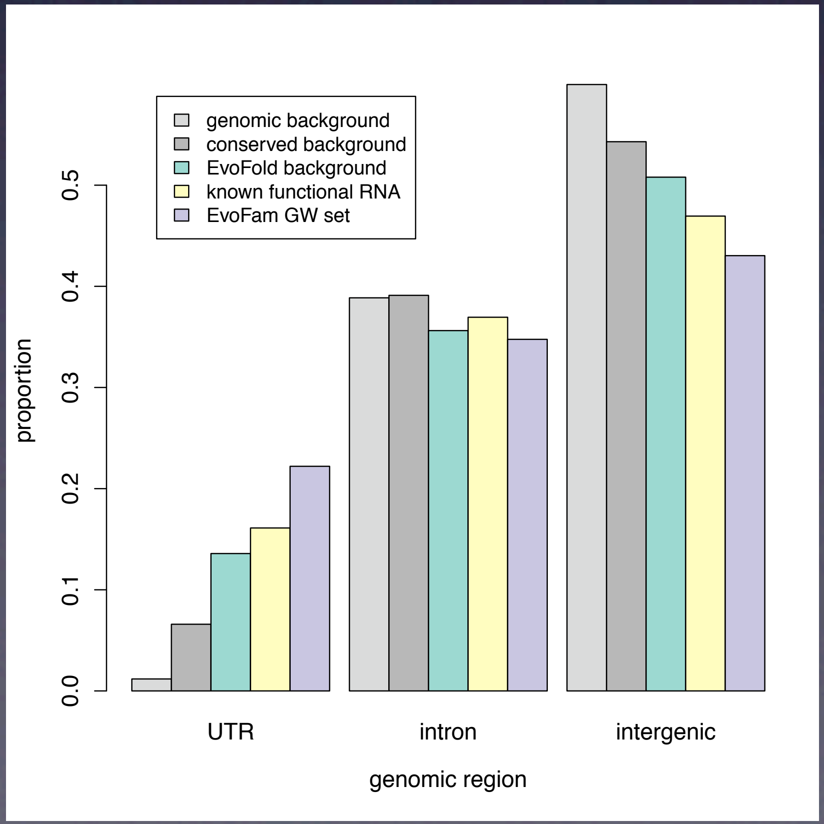


	No. of structures	No. of novel structures	No. of families	No. of novel families	EvoFold score	RNAz overlap enrichment (x)	DNAse hypersensitivity overlap (%)	Avg. correlation of tissue-specific expression within families	Intergenic expression enrichment (x)
EvoFold all (no CDS)	27,012	26,643	n/a	n/a	14	13.5	25 ($P \leq$ 5e–3)	n/a	1.20 ($P \leq$ 1e–3)
Unfiltered families	3293	3081	1254	1192	18	17.3	25 ($P \leq$ 7e–3)	0.14 ($P \leq$ 1e–3)	1.46 ($P \leq$ 1e–3)
Filtered families	725	526	220	172	18	29.0	32 ($P \leq$ 4e–3)	0.17 ($P \leq$ 1e–3)	2.33 ($P \leq$ 1e–3)

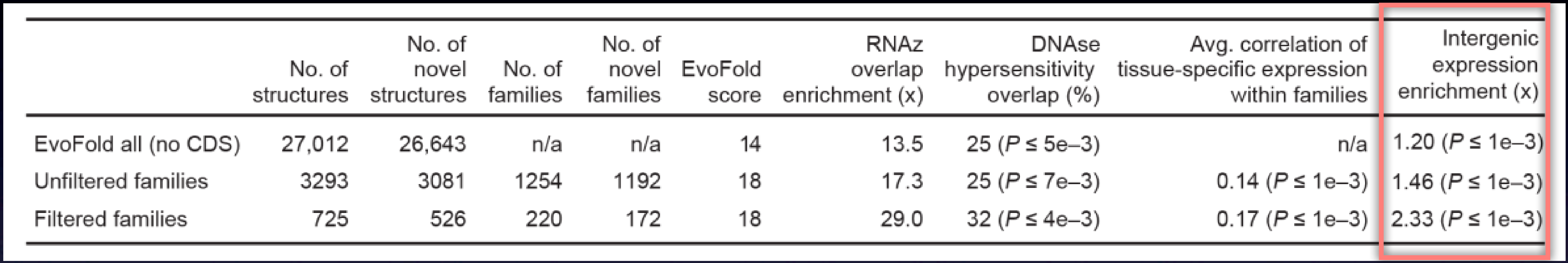
Filtered families have either:

- EvoP test < 0.05
- Region enrichment test < 0.005
- GO enrichment test relative to EvoFold background < 0.01
- Mean structure length > 11 base pairs

Genomic distribution



Family prediction overview

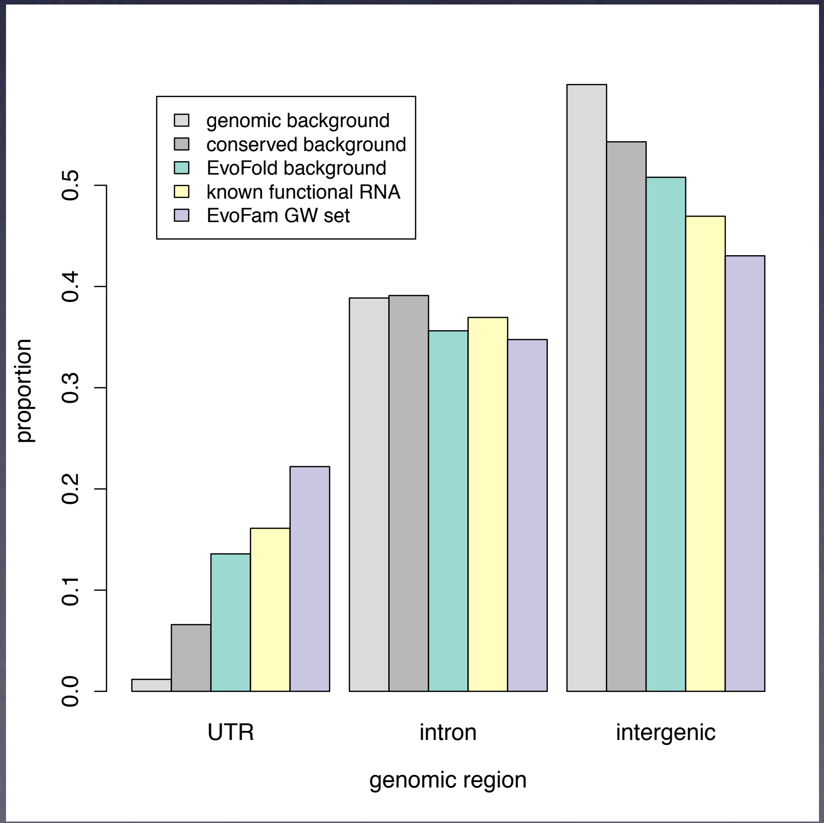


	No. of structures	No. of novel structures	No. of families	No. of novel families	EvoFold score	RNAz overlap enrichment (x)	DNAse hypersensitivity overlap (%)	Avg. correlation of tissue-specific expression within families	Intergenic expression enrichment (x)
EvoFold all (no CDS)	27,012	26,643	n/a	n/a	14	13.5	25 ($P \leq$ 5e–3)	n/a	1.20 ($P \leq$ 1e–3)
Unfiltered families	3293	3081	1254	1192	18	17.3	25 ($P \leq$ 7e–3)	0.14 ($P \leq$ 1e–3)	1.46 ($P \leq$ 1e–3)
Filtered families	725	526	220	172	18	29.0	32 ($P \leq$ 4e–3)	0.17 ($P \leq$ 1e–3)	2.33 ($P \leq$ 1e–3)

Filtered families have either:

- EvoP test < 0.05
- Region enrichment test < 0.005
- GO enrichment test relative to EvoFold background < 0.01
- Mean structure length > 11 base pairs

Genomic distribution

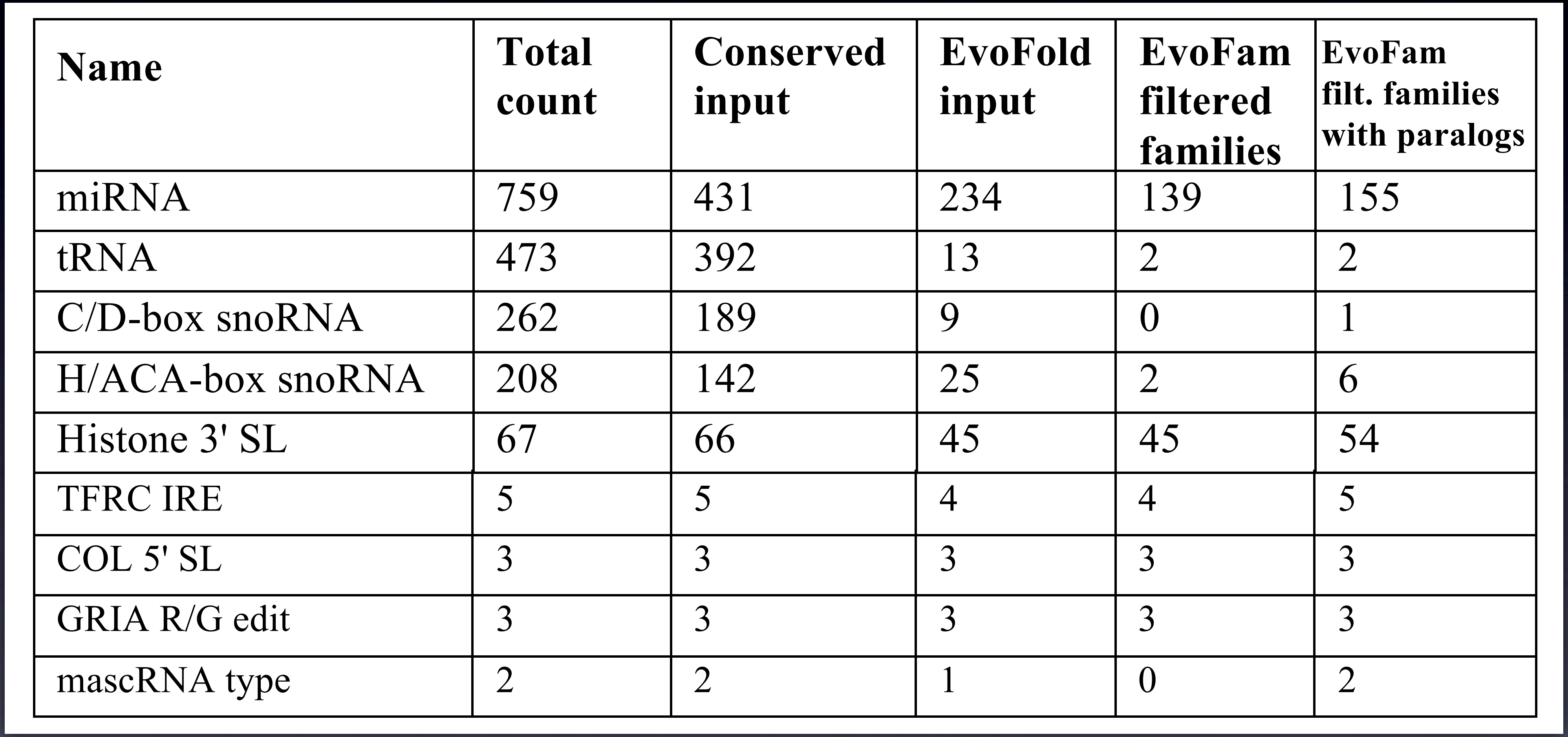


Performance

48 of 220 families contain known members (88% known)

Performance

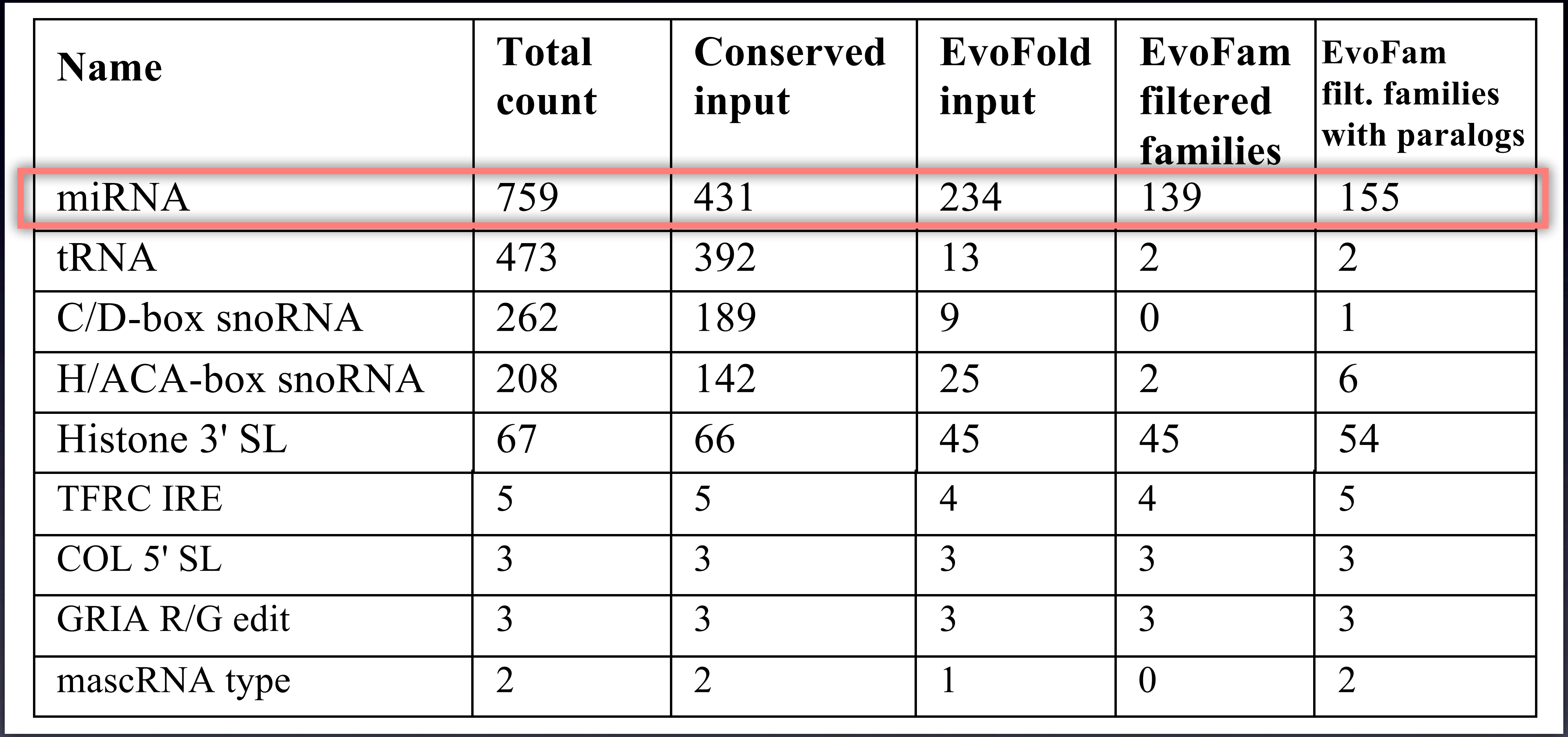
48 of 220 families contain known members (88% known)



Name	Total count	Conserved input	EvoFold input	EvoFam filtered families	EvoFam filt. families with paralogs
miRNA	759	431	234	139	155
tRNA	473	392	13	2	2
C/D-box snoRNA	262	189	9	0	1
H/ACA-box snoRNA	208	142	25	2	6
Histone 3' SL	67	66	45	45	54
TFRC IRE	5	5	4	4	5
COL 5' SL	3	3	3	3	3
GRIA R/G edit	3	3	3	3	3
mascRNA type	2	2	1	0	2

Performance

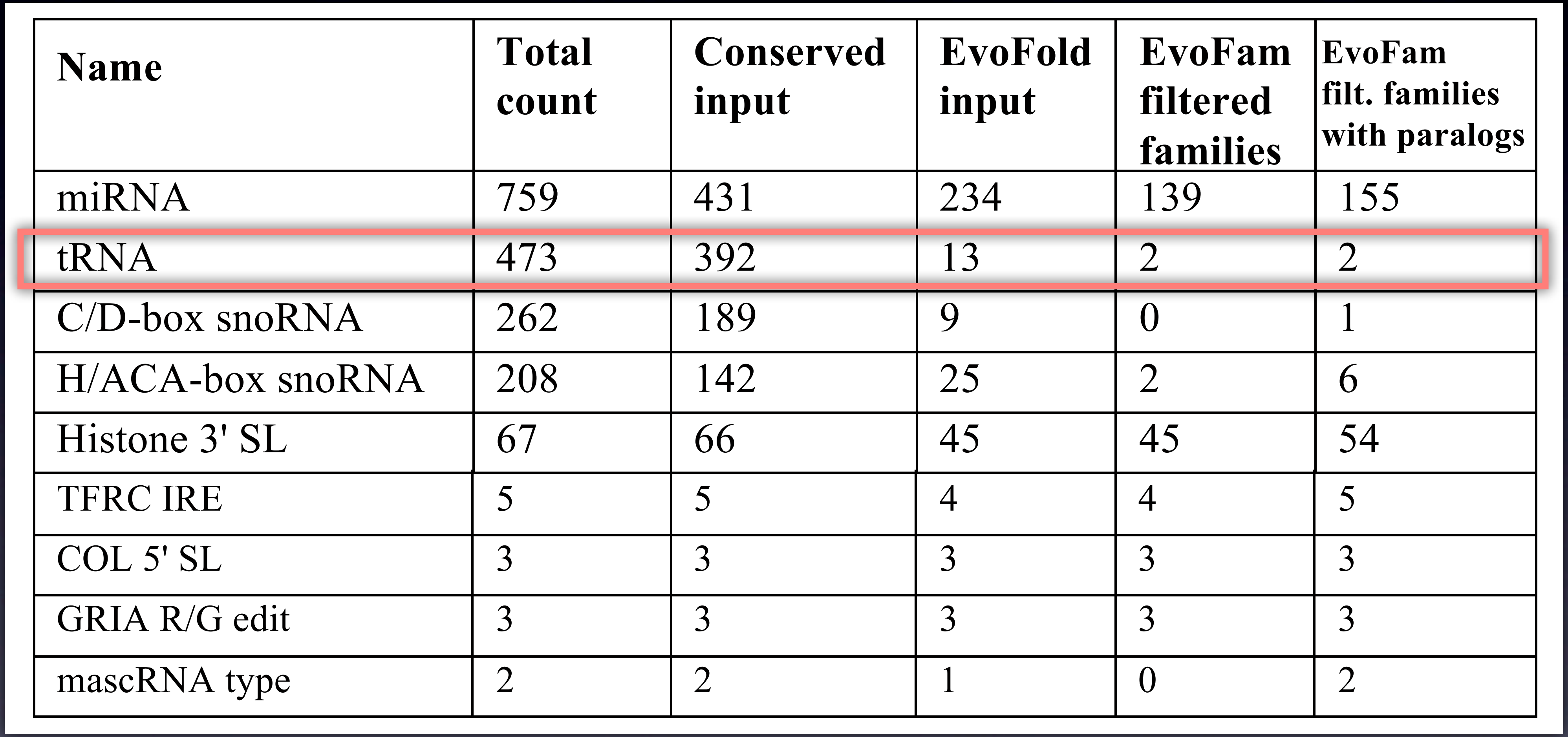
48 of 220 families contain known members (88% known)



Name	Total count	Conserved input	EvoFold input	EvoFam filtered families	EvoFam filt. families with paralogs
miRNA	759	431	234	139	155
tRNA	473	392	13	2	2
C/D-box snoRNA	262	189	9	0	1
H/ACA-box snoRNA	208	142	25	2	6
Histone 3' SL	67	66	45	45	54
TFRC IRE	5	5	4	4	5
COL 5' SL	3	3	3	3	3
GRIA R/G edit	3	3	3	3	3
mascRNA type	2	2	1	0	2

Performance

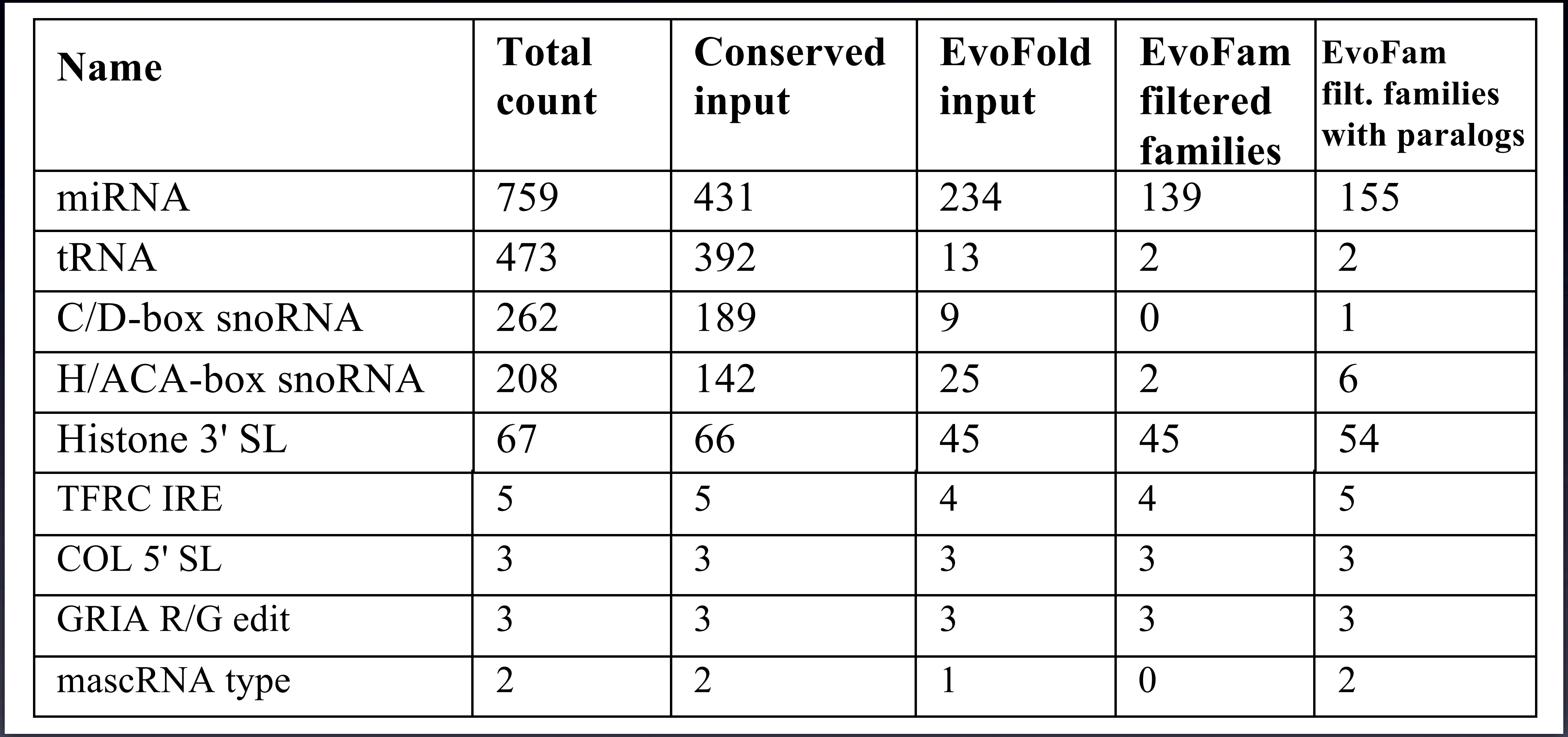
48 of 220 families contain known members (88% known)



Name	Total count	Conserved input	EvoFold input	EvoFam filtered families	EvoFam filt. families with paralogs
miRNA	759	431	234	139	155
tRNA	473	392	13	2	2
C/D-box snoRNA	262	189	9	0	1
H/ACA-box snoRNA	208	142	25	2	6
Histone 3' SL	67	66	45	45	54
TFRC IRE	5	5	4	4	5
COL 5' SL	3	3	3	3	3
GRIA R/G edit	3	3	3	3	3
mascRNA type	2	2	1	0	2

Performance

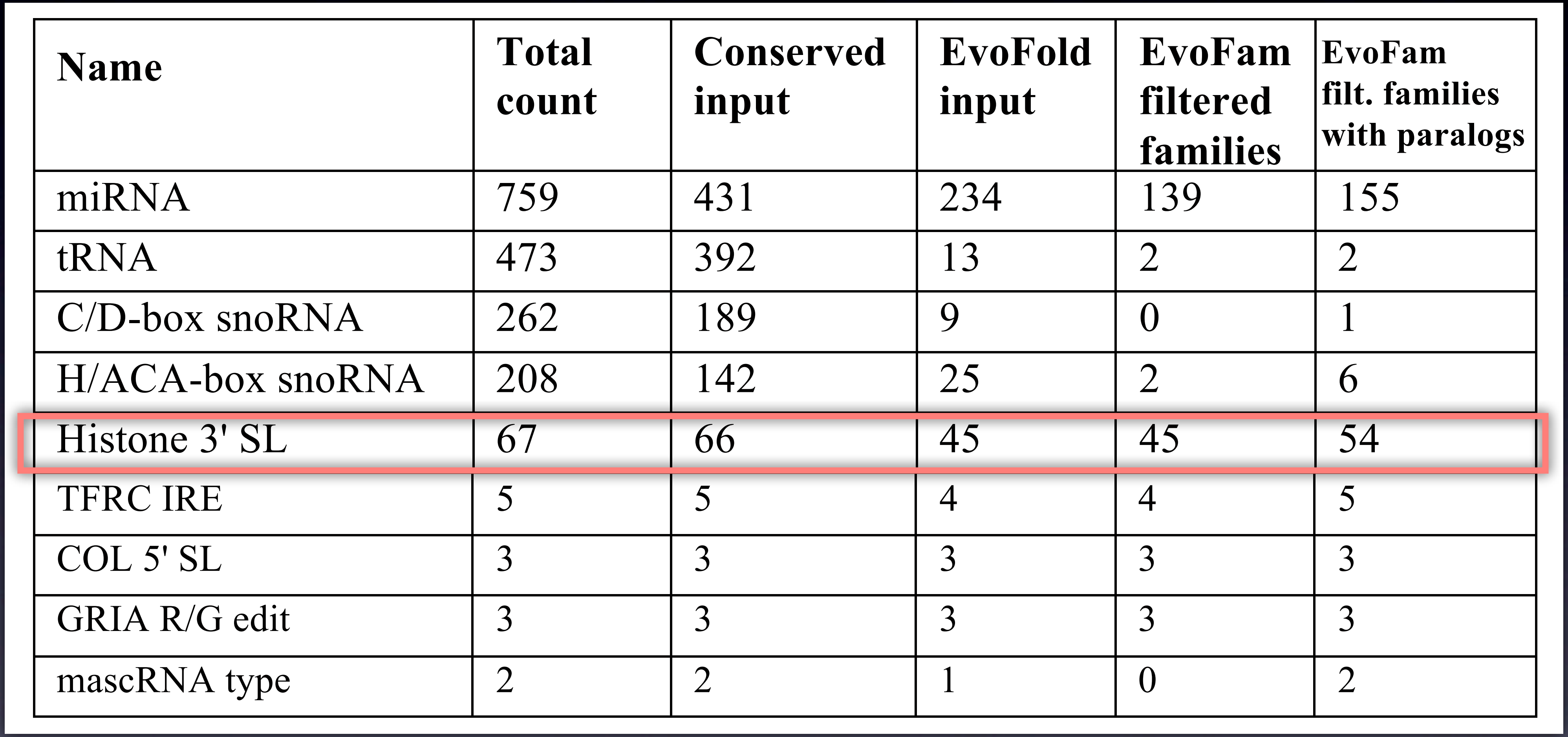
48 of 220 families contain known members (88% known)



Name	Total count	Conserved input	EvoFold input	EvoFam filtered families	EvoFam filt. families with paralogs
miRNA	759	431	234	139	155
tRNA	473	392	13	2	2
C/D-box snoRNA	262	189	9	0	1
H/ACA-box snoRNA	208	142	25	2	6
Histone 3' SL	67	66	45	45	54
TFRC IRE	5	5	4	4	5
COL 5' SL	3	3	3	3	3
GRIA R/G edit	3	3	3	3	3
mascRNA type	2	2	1	0	2

Performance

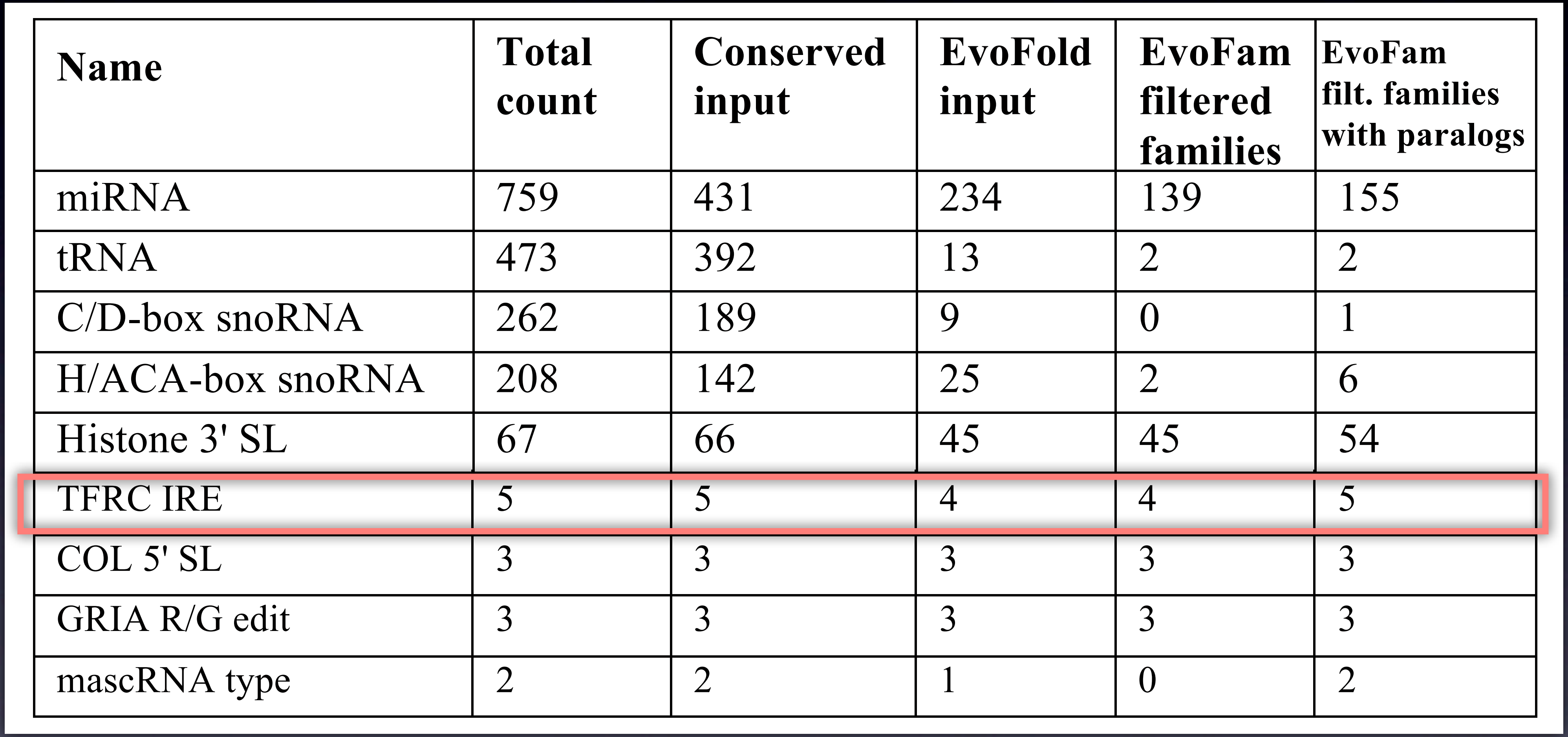
48 of 220 families contain known members (88% known)



Name	Total count	Conserved input	EvoFold input	EvoFam filtered families	EvoFam filt. families with paralogs
miRNA	759	431	234	139	155
tRNA	473	392	13	2	2
C/D-box snoRNA	262	189	9	0	1
H/ACA-box snoRNA	208	142	25	2	6
Histone 3' SL	67	66	45	45	54
TFRC IRE	5	5	4	4	5
COL 5' SL	3	3	3	3	3
GRIA R/G edit	3	3	3	3	3
mascRNA type	2	2	1	0	2

Performance

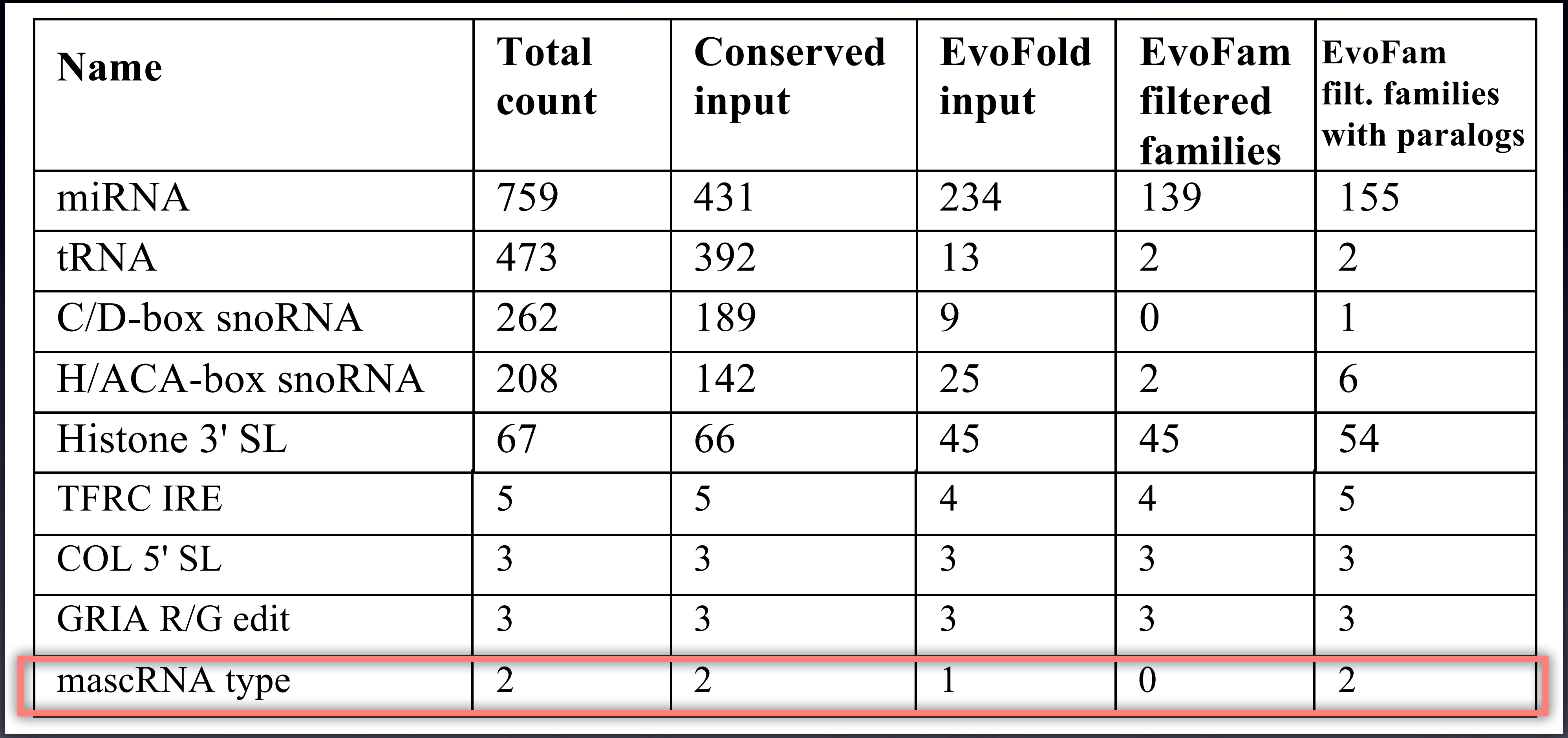
48 of 220 families contain known members (88% known)



Name	Total count	Conserved input	EvoFold input	EvoFam filtered families	EvoFam filt. families with paralogs
miRNA	759	431	234	139	155
tRNA	473	392	13	2	2
C/D-box snoRNA	262	189	9	0	1
H/ACA-box snoRNA	208	142	25	2	6
Histone 3' SL	67	66	45	45	54
TFRC IRE	5	5	4	4	5
COL 5' SL	3	3	3	3	3
GRIA R/G edit	3	3	3	3	3
mascRNA type	2	2	1	0	2

Performance

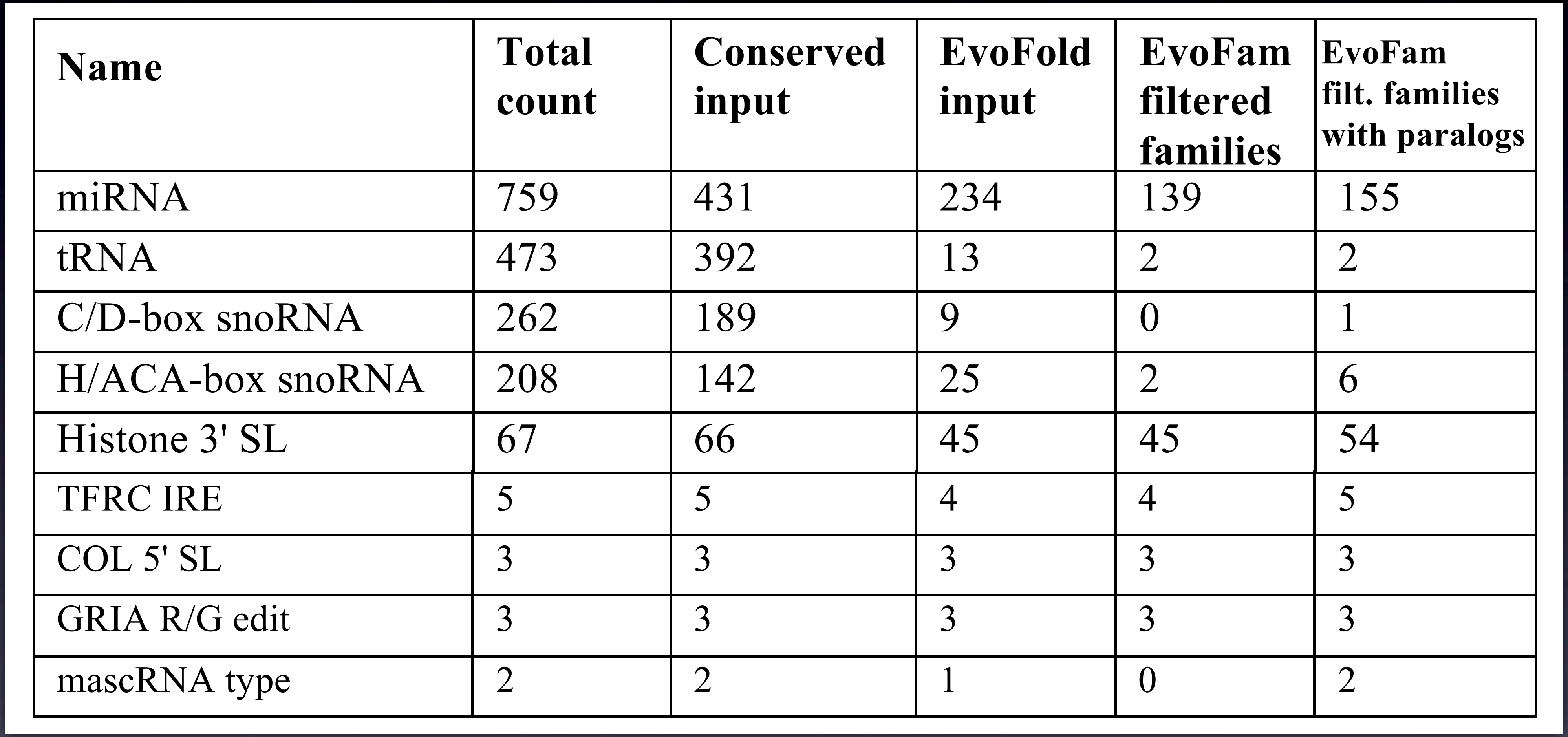
48 of 220 families contain known members (88% known)



Name	Total count	Conserved input	EvoFold input	EvoFam filtered families	EvoFam filt. families with paralogs
miRNA	759	431	234	139	155
tRNA	473	392	13	2	2
C/D-box snoRNA	262	189	9	0	1
H/ACA-box snoRNA	208	142	25	2	6
Histone 3' SL	67	66	45	45	54
TFRC IRE	5	5	4	4	5
COL 5' SL	3	3	3	3	3
GRIA R/G edit	3	3	3	3	3
mascRNA type	2	2	1	0	2

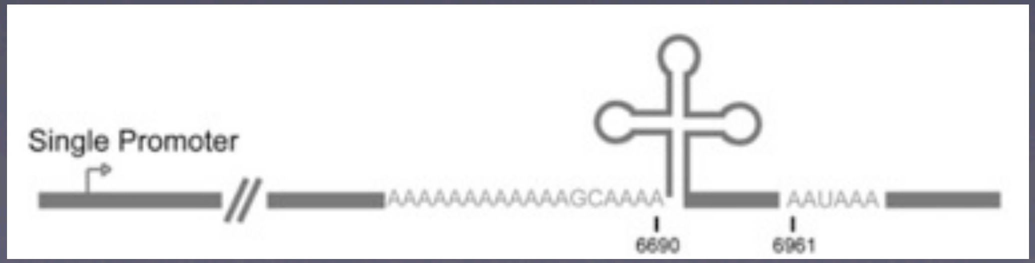
Performance

48 of 220 families contain known members (88% known)



Name	Total count	Conserved input	EvoFold input	EvoFam filtered families	EvoFam filt. families with paralogs
miRNA	759	431	234	139	155
tRNA	473	392	13	2	2
C/D-box snoRNA	262	189	9	0	1
H/ACA-box snoRNA	208	142	25	2	6
Histone 3' SL	67	66	45	45	54
TFRC IRE	5	5	4	4	5
COL 5' SL	3	3	3	3	3
GRIA R/G edit	3	3	3	3	3
mascRNA type	2	2	1	0	2

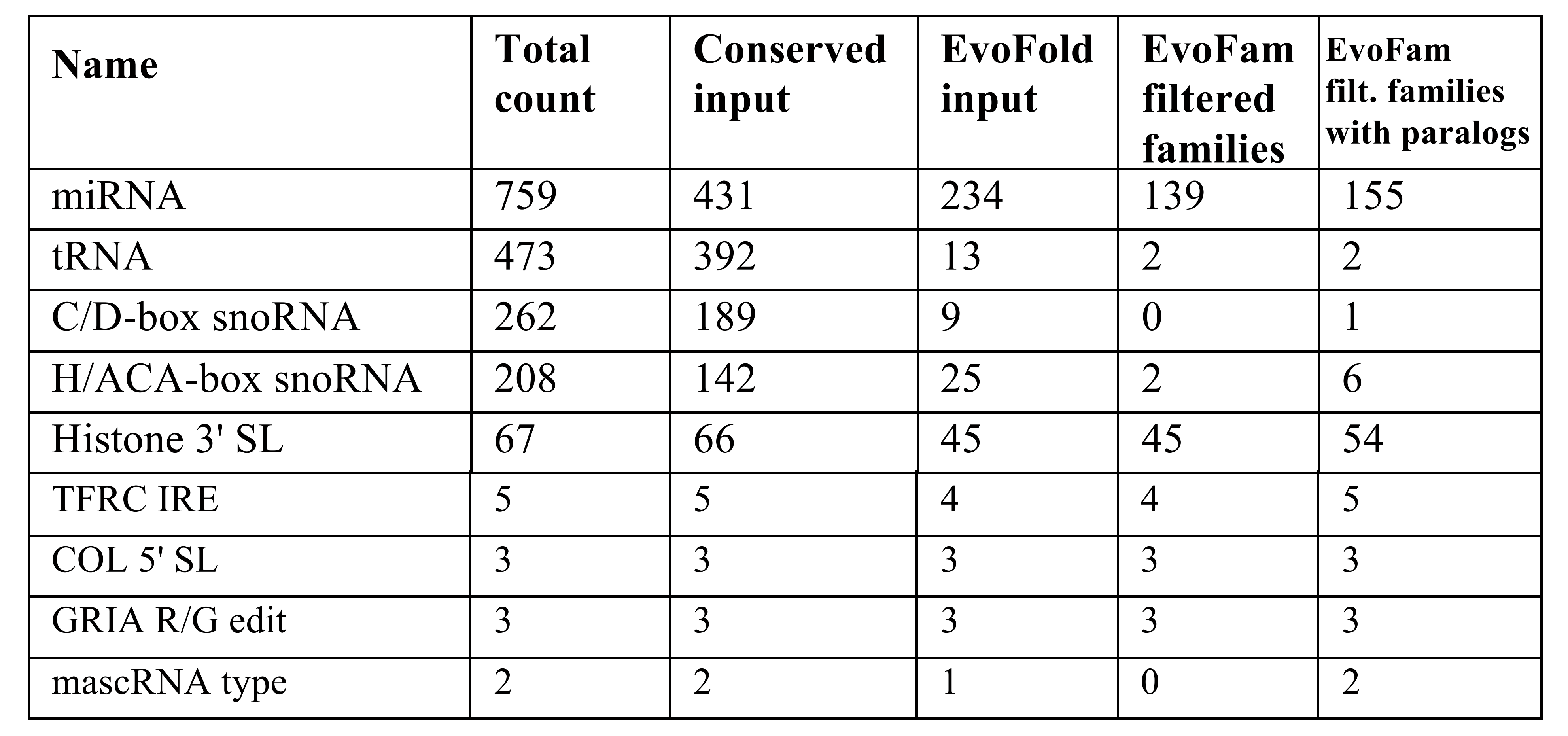
mascRNA family in MALAT 1 and Men β



Wilusz 2008; Sunwoo et al. 2009; Wilusz and Spector 2010

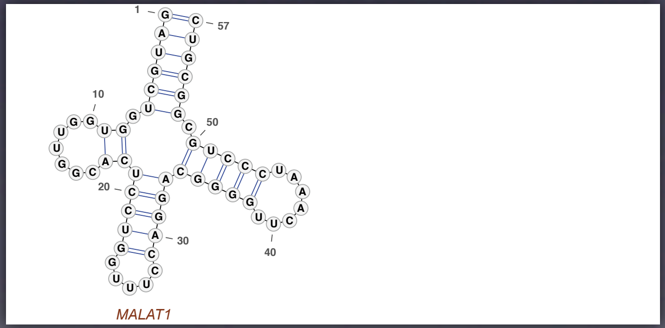
Performance

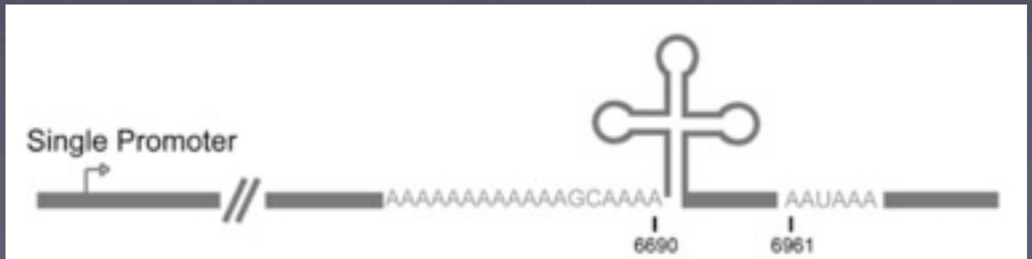
48 of 220 families contain known members (88% known)



Name	Total count	Conserved input	EvoFold input	EvoFam filtered families	EvoFam filt. families with paralogs
miRNA	759	431	234	139	155
tRNA	473	392	13	2	2
C/D-box snoRNA	262	189	9	0	1
H/ACA-box snoRNA	208	142	25	2	6
Histone 3' SL	67	66	45	45	54
TFRC IRE	5	5	4	4	5
COL 5' SL	3	3	3	3	3
GRIA R/G edit	3	3	3	3	3
mascRNA type	2	2	1	0	2

mascRNA family in MALAT 1 and Men β

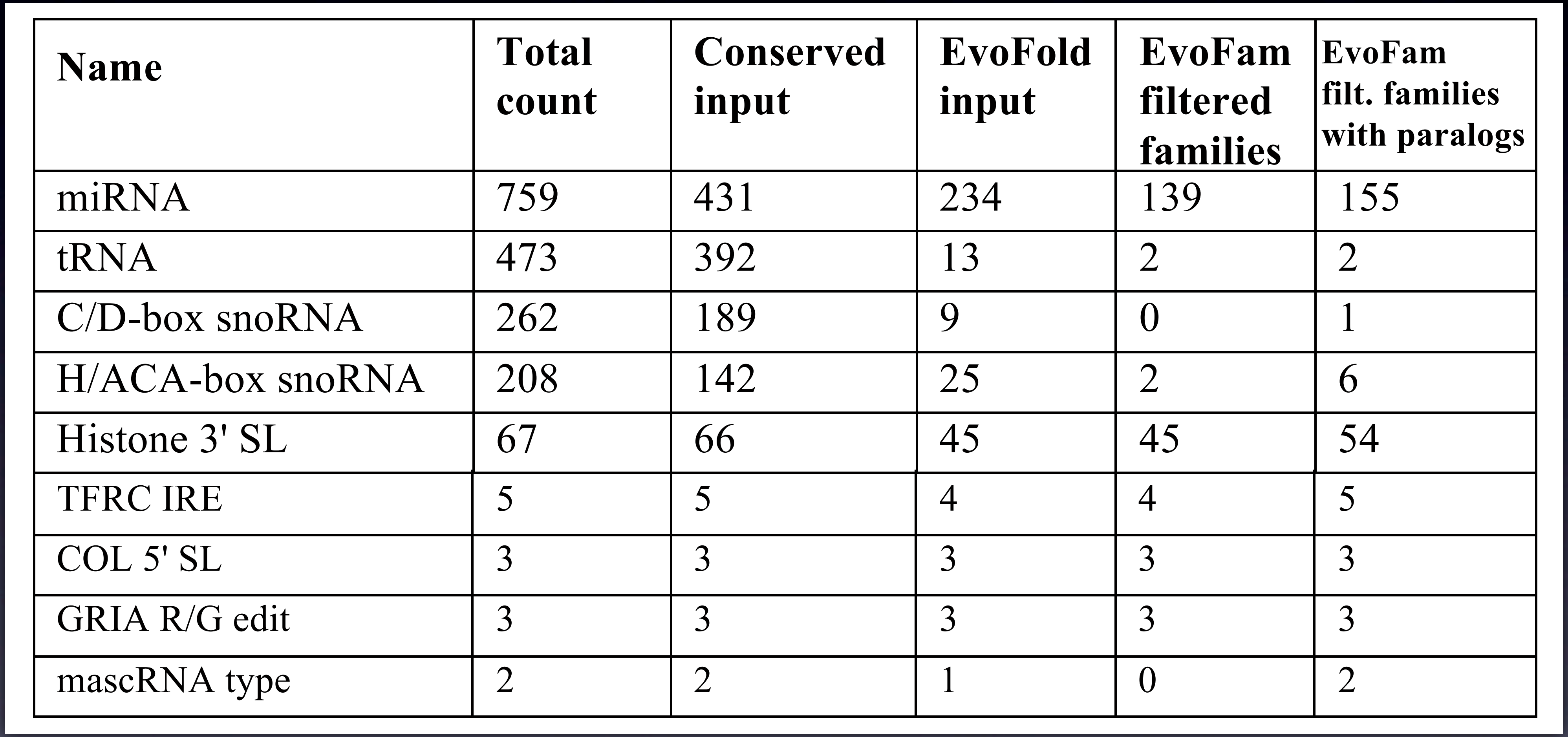




Wilusz 2008; Sunwoo et al. 2009; Wilusz and Spector 2010

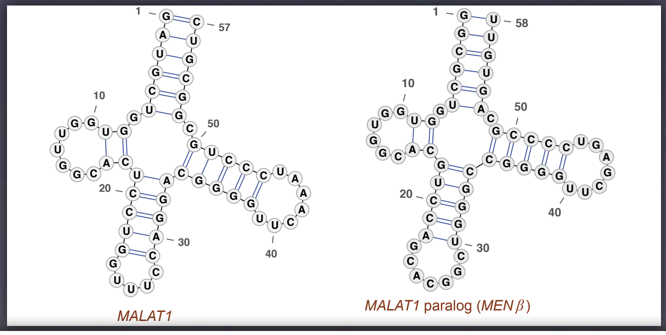
Performance

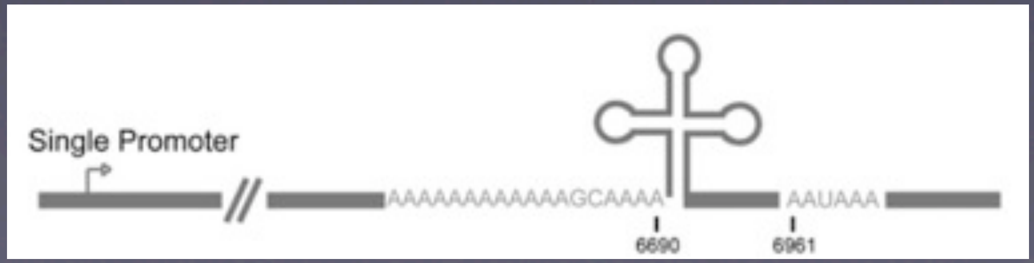
48 of 220 families contain known members (88% known)

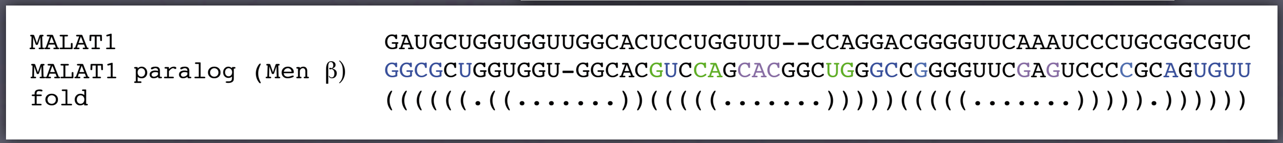


Name	Total count	Conserved input	EvoFold input	EvoFam filtered families	EvoFam filt. families with paralogs
miRNA	759	431	234	139	155
tRNA	473	392	13	2	2
C/D-box snoRNA	262	189	9	0	1
H/ACA-box snoRNA	208	142	25	2	6
Histone 3' SL	67	66	45	45	54
TFRC IRE	5	5	4	4	5
COL 5' SL	3	3	3	3	3
GRIA R/G edit	3	3	3	3	3
mascRNA type	2	2	1	0	2

mascRNA family in MALAT 1 and Men β





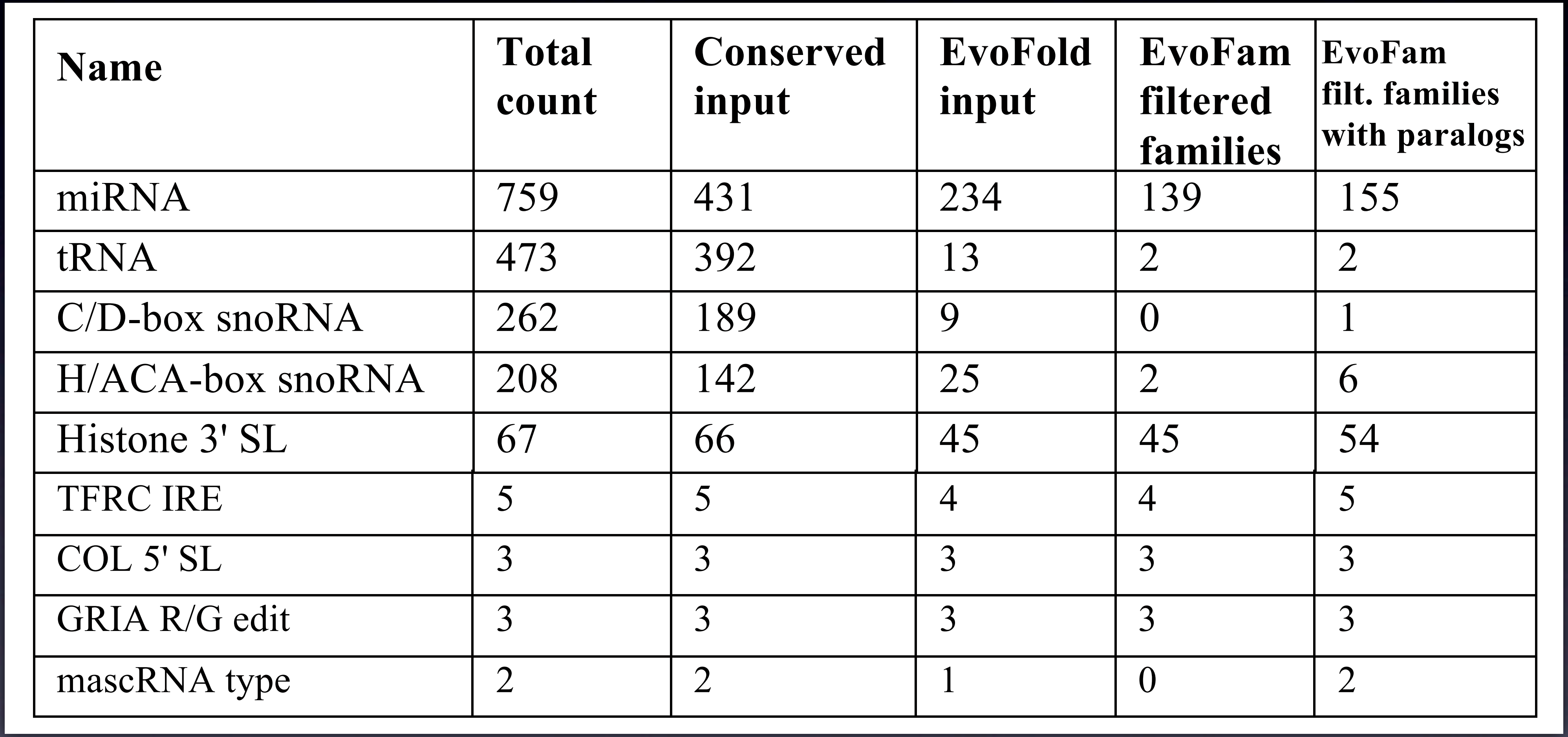


```
MALAT1                  GAUGCUGGUGGUUGGCACUCCUGGUUU--CCAGGACGGGGUUCAAAUCCCUGCGGCGUC
MALAT1 paralog (Men β)  GGCGCUGGUGGU-GGCACGUCCAGCACGGCUGGCCGGGGUUCGAGUCCCCGCAGUGUU
fold                    ((((((.((.......))(((((.......)))))(((((.......))))).))))))
```

Wilusz 2008; Sunwoo et al. 2009; Wilusz and Spector 2010

Performance

48 of 220 families contain known members (88% known)



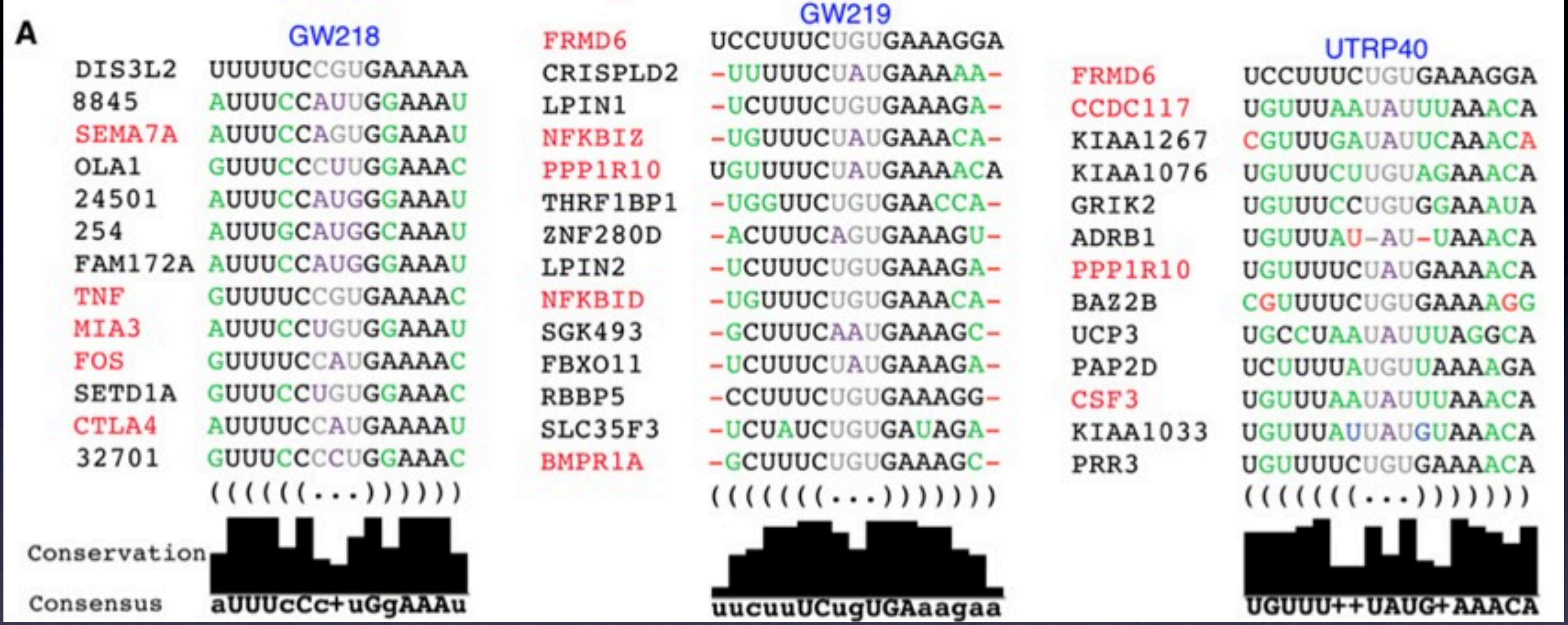
Name	Total count	Conserved input	EvoFold input	EvoFam filtered families	EvoFam filt. families with paralogs
miRNA	759	431	234	139	155
tRNA	473	392	13	2	2
C/D-box snoRNA	262	189	9	0	1
H/ACA-box snoRNA	208	142	25	2	6
Histone 3' SL	67	66	45	45	54
TFRC IRE	5	5	4	4	5
COL 5' SL	3	3	3	3	3
GRIA R/G edit	3	3	3	3	3
mascRNA type	2	2	1	0	2

Estimated false positive rate:

- 27 % for similarity edges
- 34 % for families of size three or larger

Immune related regulatory networks?

Families of short hairpins enriched in 3‘UTRs of immunity related genes



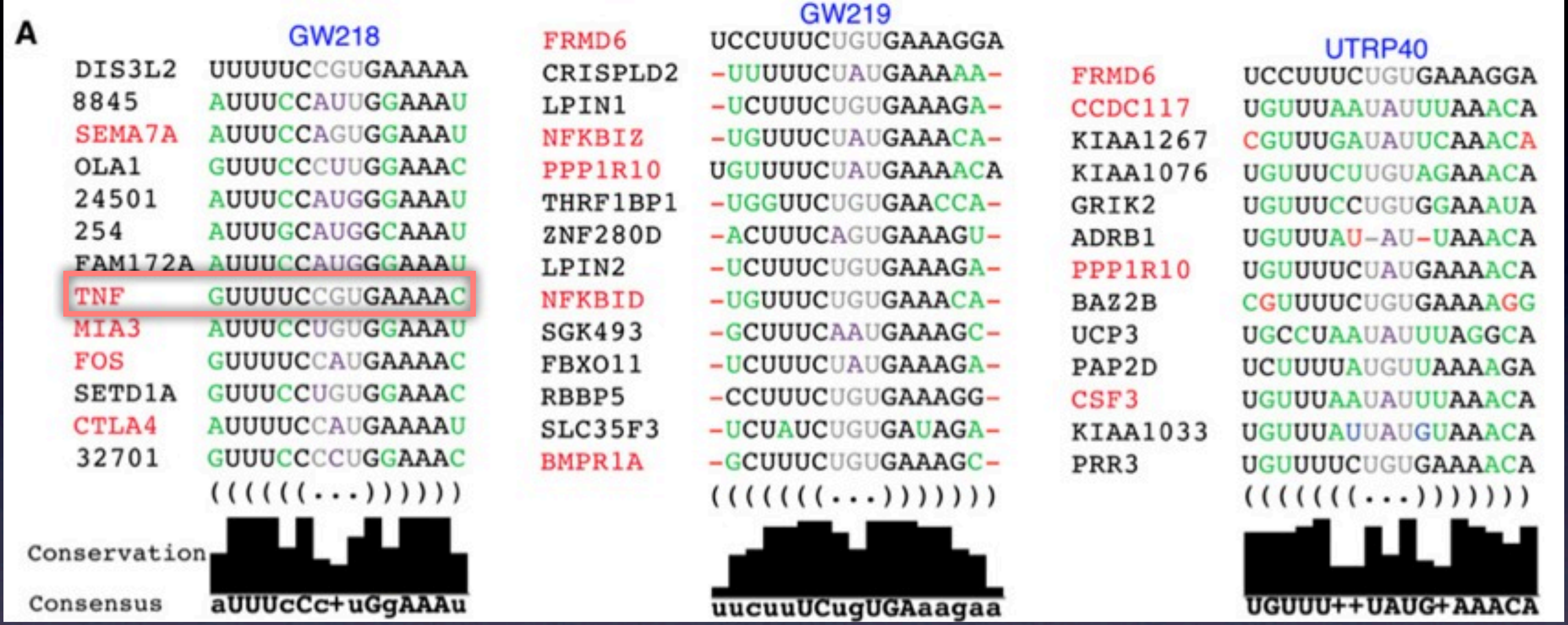
	GW218
DIS3L2	UUUUUCCGUGAAAAA
8845	AUUUCCAUUGGAAAU
SEMA7A	AUUUCCAGUGGAAAU
OLA1	GUUUCCCUUGGAAAC
24501	AUUUCCAUGGGAAAU
254	AUUUGCAUGGCAAAU
FAM172A	AUUUCCAUGGGAAAU
TNF	GUUUUCCGUGAAAAC
MIA3	AUUUCCUGUGGAAAU
FOS	GUUUUCCAUGAAAAC
SETD1A	GUUUCCUGUGGAAAC
CTLA4	AUUUUCCAUGAAAAU
32701	GUUUCCCCUGGAAAC
	((((((...))))))
Consensus	aUUUcCc+uGgAAAu

	GW219
FRMD6	UCCUUUCUGUGAAAGGA
CRISPLD2	-UUUUUCUAUGAAAAA-
LPIN1	-UCUUUCUGUGAAAGA-
NFKBIZ	-UGUUUCUAUGAAACA-
PPP1R10	UGUUUUCUAUGAAAACA
THRF1BP1	-UGGUUCUGUGAACCA-
ZNF280D	-ACUUUCAGUGAAAGU-
LPIN2	-UCUUUCUGUGAAAGA-
NFKBID	-UGUUUCUGUGAAACA-
SGK493	-GCUUUCAAUGAAAGC-
FBXO11	-UCUUUCUAUGAAAGA-
RBBP5	-CCUUUCUGUGAAAGG-
SLC35F3	-UCUAUCUGUGAUAGA-
BMPR1A	-GCUUUCUGUGAAAGC-
	(((((((...)))))))
	uucuuUCugUGAaagaa

	UTRP40
FRMD6	UCCUUUCUGUGAAAGGA
CCDC117	UGUUUAAUAUUUAAACA
KIAA1267	CGUUUGAUAUUCAAACA
KIAA1076	UGUUUCUUGUAGAAACA
GRIK2	UGUUUCCUGUGGAAAUA
ADRB1	UGUUUAU-AU-UAAACA
PPP1R10	UGUUUUCUAUGAAAACA
BAZ2B	CGUUUUCUGUGAAAAGG
UCP3	UGCCUAAUAUUUAGGCA
PAP2D	UCUUUUAUGUUAAAAGA
CSF3	UGUUUAAUAUUUAAACA
KIAA1033	UGUUUAUUAUGUAAACA
PRR3	UGUUUUCUGUGAAAACA
	(((((((...)))))))
	UGUUU++UAUG+AAACA

Immune related regulatory networks?

Families of short hairpins enriched in 3‘UTRs of immunity related genes


A
Conservation

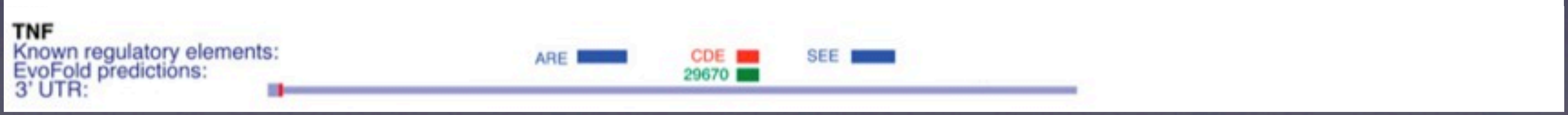
	GW218
DIS3L2	UUUUUCCGUGAAAAA
8845	AUUUCCAUUGGAAAU
SEMA7A	AUUUCCAGUGGAAAU
OLA1	GUUUCCCUUGGAAAC
24501	AUUUCCAUGGGAAAU
254	AUUUGCAUGGCAAAU
FAM172A	AUUUCCAUGGGAAAU
TNF	GUUUUCCGUGAAAAC
MIA3	AUUUCCUGUGGAAAU
FOS	GUUUUCCAUGAAAAC
SETD1A	GUUUCCUGUGGAAAC
CTLA4	AUUUUCCAUGAAAAU
32701	GUUUCCCCUGGAAAC
	((((((...))))))
Consensus	aUUUcCc+uGgAAAu

	GW219
FRMD6	UCCUUUCUGUGAAAGGA
CRISPLD2	-UUUUUCUAUGAAAAA-
LPIN1	-UCUUUCUGUGAAAGA-
NFKBIZ	-UGUUUCUAUGAAACA-
PPP1R10	UGUUUUCUAUGAAAACA
THRF1BP1	-UGGUUCUGUGAACCA-
ZNF280D	-ACUUUCAGUGAAAGU-
LPIN2	-UCUUUCUGUGAAAGA-
NFKBID	-UGUUUCUGUGAAACA-
SGK493	-GCUUUCAAUGAAAGC-
FBXO11	-UCUUUCUAUGAAAGA-
RBBP5	-CCUUUCUGUGAAAGG-
SLC35F3	-UCUAUCUGUGAUAGA-
BMPR1A	-GCUUUCUGUGAAAGC-
	(((((((...)))))))
	uucuuUCugUGAaagaa

	UTRP40
FRMD6	UCCUUUCUGUGAAAGGA
CCDC117	UGUUUAAUAUUUAAACA
KIAA1267	CGUUUGAUAUUCAAACA
KIAA1076	UGUUUCUUGUAGAAACA
GRIK2	UGUUUCCUGUGGAAAUA
ADRB1	UGUUUAU-AU-UAAACA
PPP1R10	UGUUUUCUAUGAAAACA
BAZ2B	CGUUUUCUGUGAAAAGG
UCP3	UGCCUAAUAUUUAGGCA
PAP2D	UCUUUUAUGUUAAAAGA
CSF3	UGUUUAAUAUUUAAACA
KIAA1033	UGUUUAUUAUGUAAACA
PRR3	UGUUUUCUGUGAAAACA
	(((((((...)))))))
	UGUUU++UAUG+AAACA

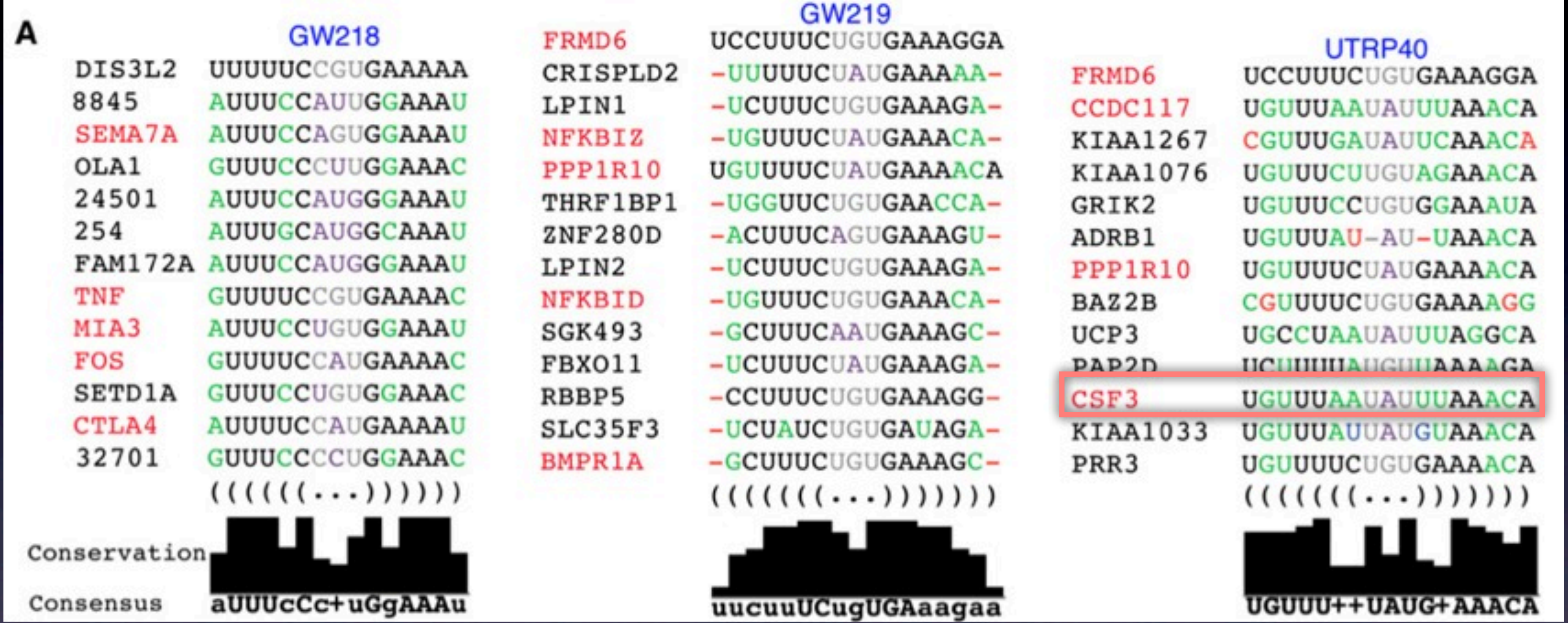
Includes known destabilization hairpins

Constitutive Decay Element



Immune related regulatory networks?

Families of short hairpins enriched in 3‘UTRs of immunity related genes



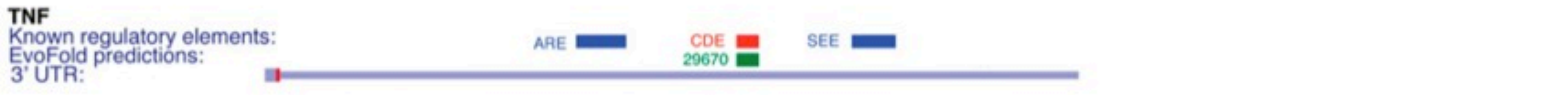
	GW218
DIS3L2	UUUUUCCGUGAAAAA
8845	AUUUCCAUUGGAAAU
SEMA7A	AUUUCCAGUGGAAAU
OLA1	GUUUCCCUUGGAAAC
24501	AUUUCCAUGGGAAAU
254	AUUUGCAUGGCAAAU
FAM172A	AUUUCCAUGGGAAAU
TNF	GUUUUCCGUGAAAAC
MIA3	AUUUCCUGUGGAAAU
FOS	GUUUUCCAUGAAAAC
SETD1A	GUUUCCUGUGGAAAC
CTLA4	AUUUUCCAUGAAAAU
32701	GUUUCCCCUGGAAAC
	((((((...))))))
Conservation	
Consensus	aUUUcCc+uGgAAAu

	GW219
FRMD6	UCCUUUCUGUGAAAGGA
CRISPLD2	-UUUUUCUAUGAAAAA-
LPIN1	-UCUUUCUGUGAAAGA-
NFKBIZ	-UGUUUCUAUGAAACA-
PPP1R10	UGUUUUCUAUGAAAACA
THRF1BP1	-UGGUUCUGUGAACCA-
ZNF280D	-ACUUUCAGUGAAAGU-
LPIN2	-UCUUUCUGUGAAAGA-
NFKBID	-UGUUUCUGUGAAACA-
SGK493	-GCUUUCAAUGAAAGC-
FBXO11	-UCUUUCUAUGAAAGA-
RBBP5	-CCUUUCUGUGAAAGG-
SLC35F3	-UCUAUCUGUGAUAGA-
BMPR1A	-GCUUUCUGUGAAAGC-
	(((((((...)))))))
	uucuuUCugUGAaagaa

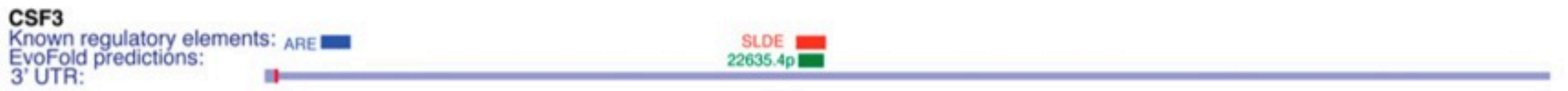
	UTRP40
FRMD6	UCCUUUCUGUGAAAGGA
CCDC117	UGUUUAAUAUUUAAACA
KIAA1267	CGUUUGAUAUUCAAACA
KIAA1076	UGUUUCUUGUAGAAACA
GRIK2	UGUUUCCUGUGGAAAUA
ADRB1	UGUUUAU-AU-UAAACA
PPP1R10	UGUUUUCUAUGAAAACA
BAZ2B	CGUUUUCUGUGAAAAGG
UCP3	UGCCUAAUAUUUAGGCA
PAP2D	UCUUUUAUGUUAAAAGA
CSF3	UGUUUAAUAUUUAAACA
KIAA1033	UGUUUAUUAUGUAAACA
PRR3	UGUUUUCUGUGAAAACA
	(((((((...)))))))
	UGUUU++UAUG+AAACA

Includes known destabilization hairpins

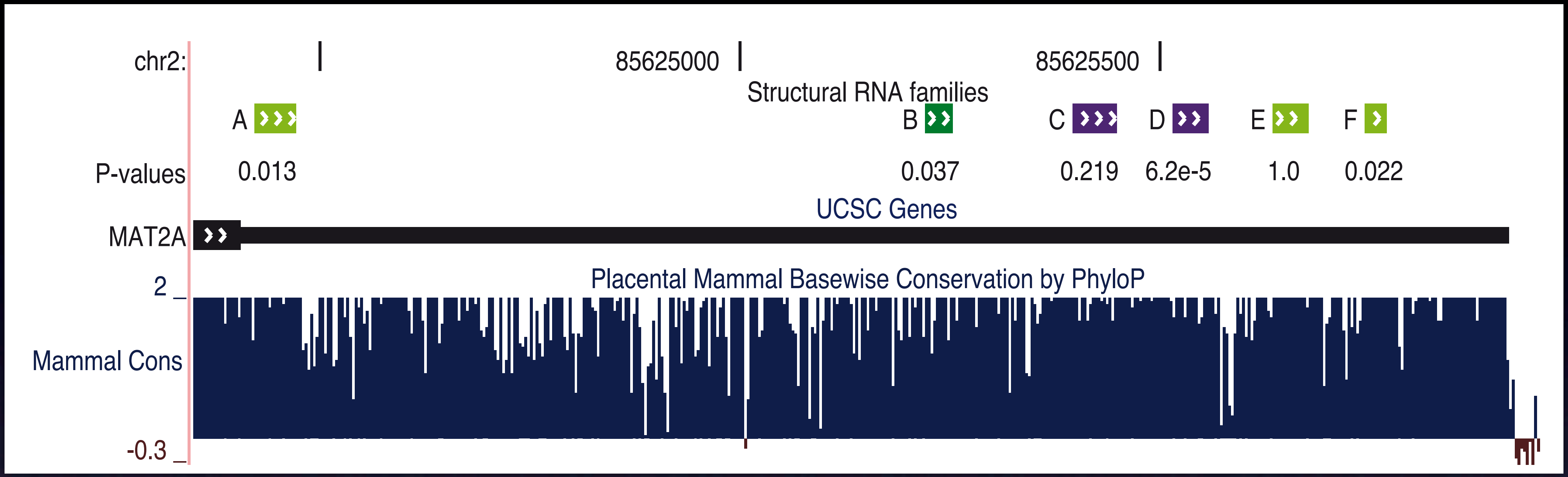
Constitutive Decay Element



Stem–Loop Destabilizing Element

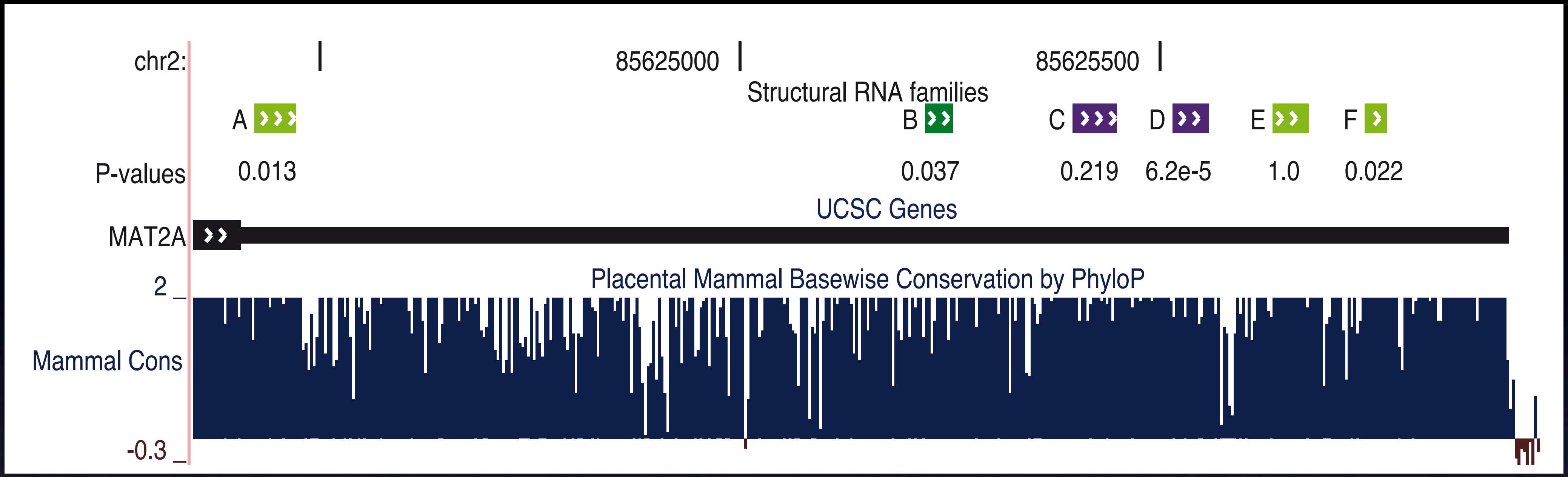


Family of six hairpins all within 3‘UTR MAT2A

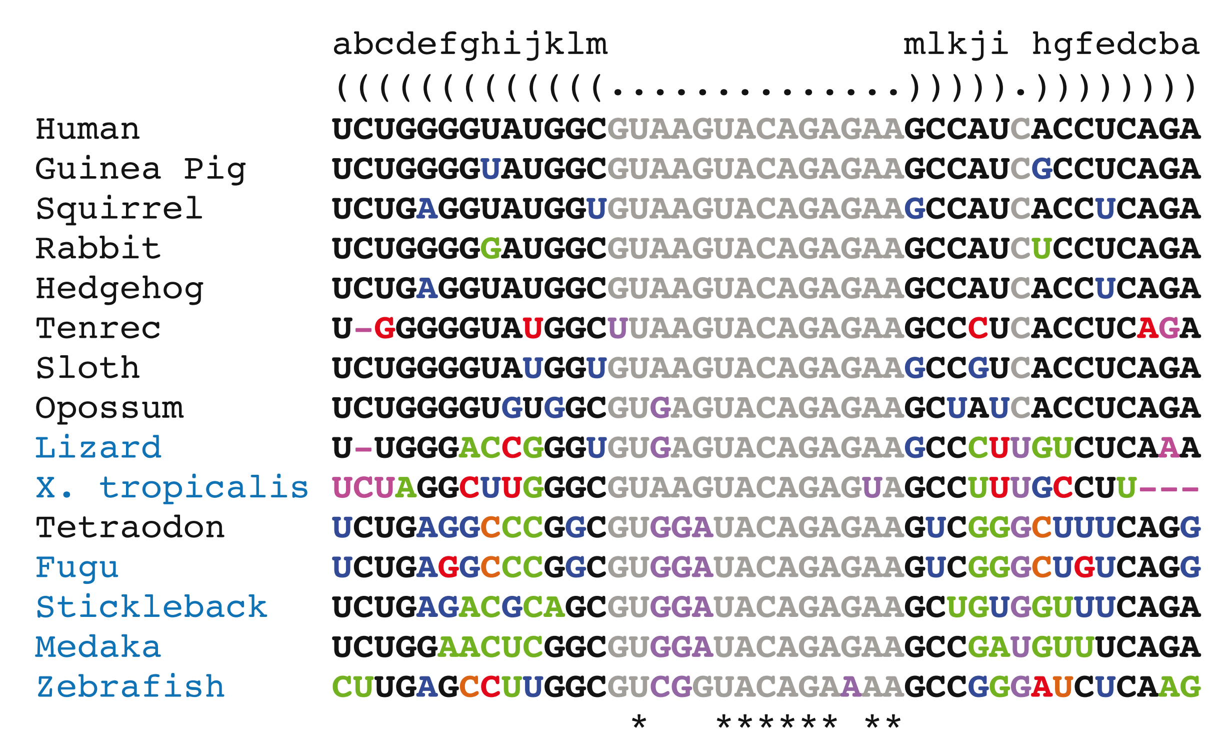


Brian J. Parker, et al. New families of human regulatory RNA structures identified by comparative analysis of vertebrate genomes. Genome Research (2011).

Family of six hairpins all within 3‘UTR MAT2A

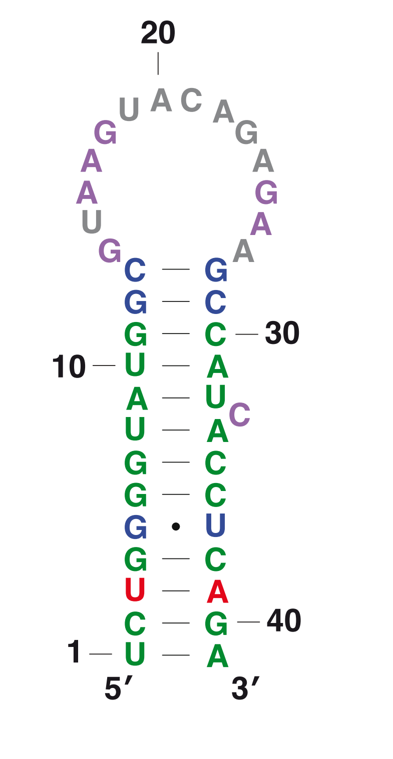


Vertebrate alignment for hairpin D



```
              abcdefghijklm              mlkji hgfedcba
              (((((((((((((..............))))).))))))))
Human         UCUGGGGUAUGGCGUAAGUACAGAGAAGCCAUCACCUCAGA
Guinea Pig    UCUGGGGUAUGGCGUAAGUACAGAGAAGCCAUCGCCUCAGA
Squirrel      UCUGAGGUAUGGUGUAAGUACAGAGAAGCCAUCACCUCAGA
Rabbit        UCUGGGGGAUGGCGUAAGUACAGAGAAGCCAUCUCCUCAGA
Hedgehog      UCUGAGGUAUGGCGUAAGUACAGAGAAGCCAUCACCUCAGA
Tenrec        U-GGGGGUAUGGCUUAAGUACAGAGAAGCCCUCACCUCAGA
Sloth         UCUGGGGUAUGGUGUAAGUACAGAGAAGCCGUCACCUCAGA
Opossum       UCUGGGGUGUGGCGUGAGUACAGAGAAGCUAUCACCUCAGA
Lizard        U-UGGGACCGGGUGUGAGUACAGAGAAGCCCUUGUCUCAAA
X. tropicalis UCUAGGCUUGGGCGUAAGUACAGAGUAGCCUUUGCCUU---
Tetraodon     UCUGAGGCCCGGCGUGGAUACAGAGAAGUCGGGCUUUCAGG
Fugu          UCUGAGGCCCGGCGUGGAUACAGAGAAGUCGGGCUGUCAGG
Stickleback   UCUGAGACGCAGCGUGGAUACAGAGAAGCUGUGGUUUCAGA
Medaka        UCUGGAACUCGGCGUGGAUACAGAGAAGCCGAUGUUUCAGA
Zebrafish     CUUGAGCCUUGGCGUCGGUACAGAAAAGCCGGGAUCUCAAG
                            *   ****** **
```

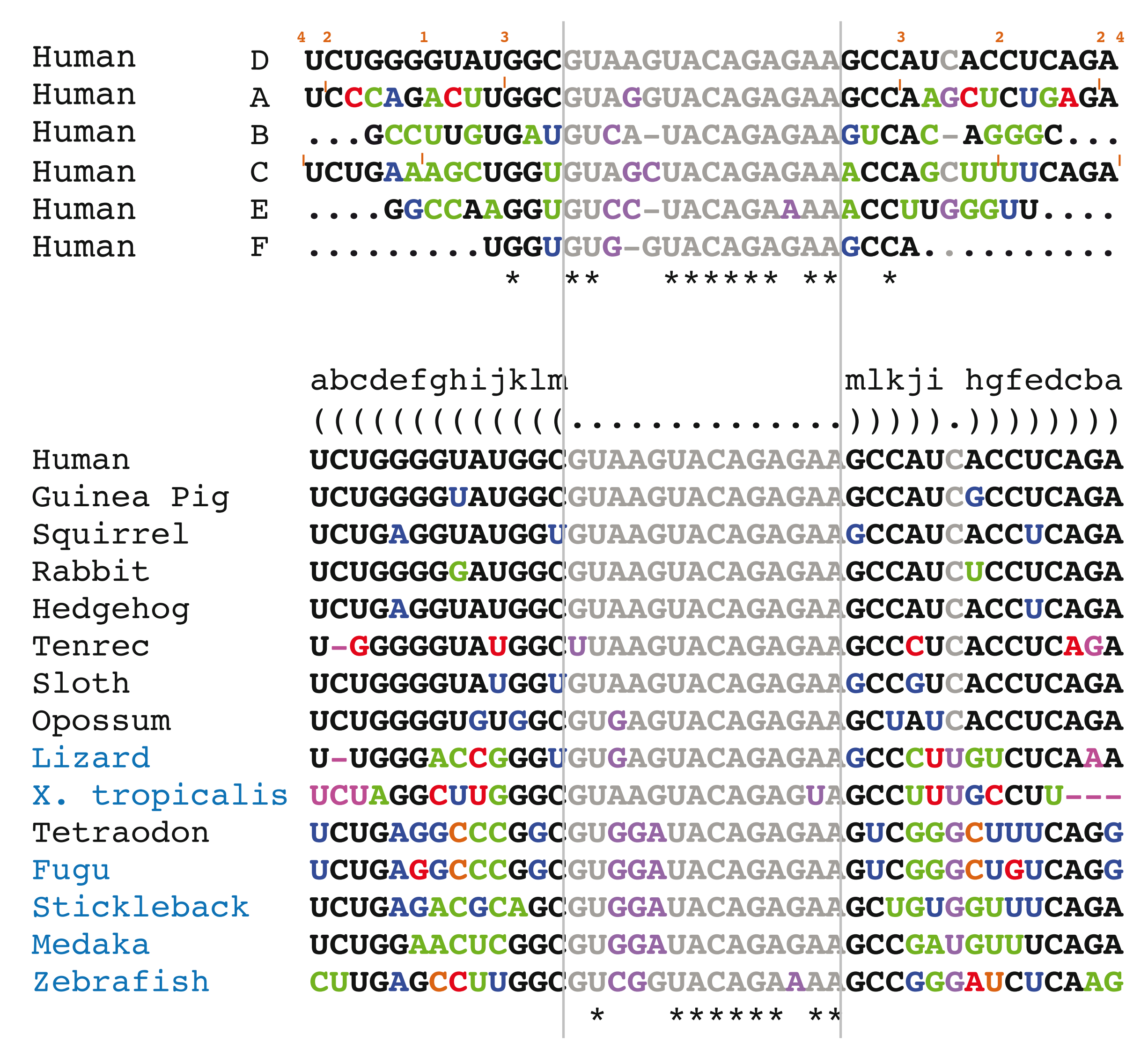
Hairpin D



Brian J. Parker, et al. New families of human regulatory RNA structures identified by comparative analysis of vertebrate genomes. Genome Research (2011).

Shared loop motif

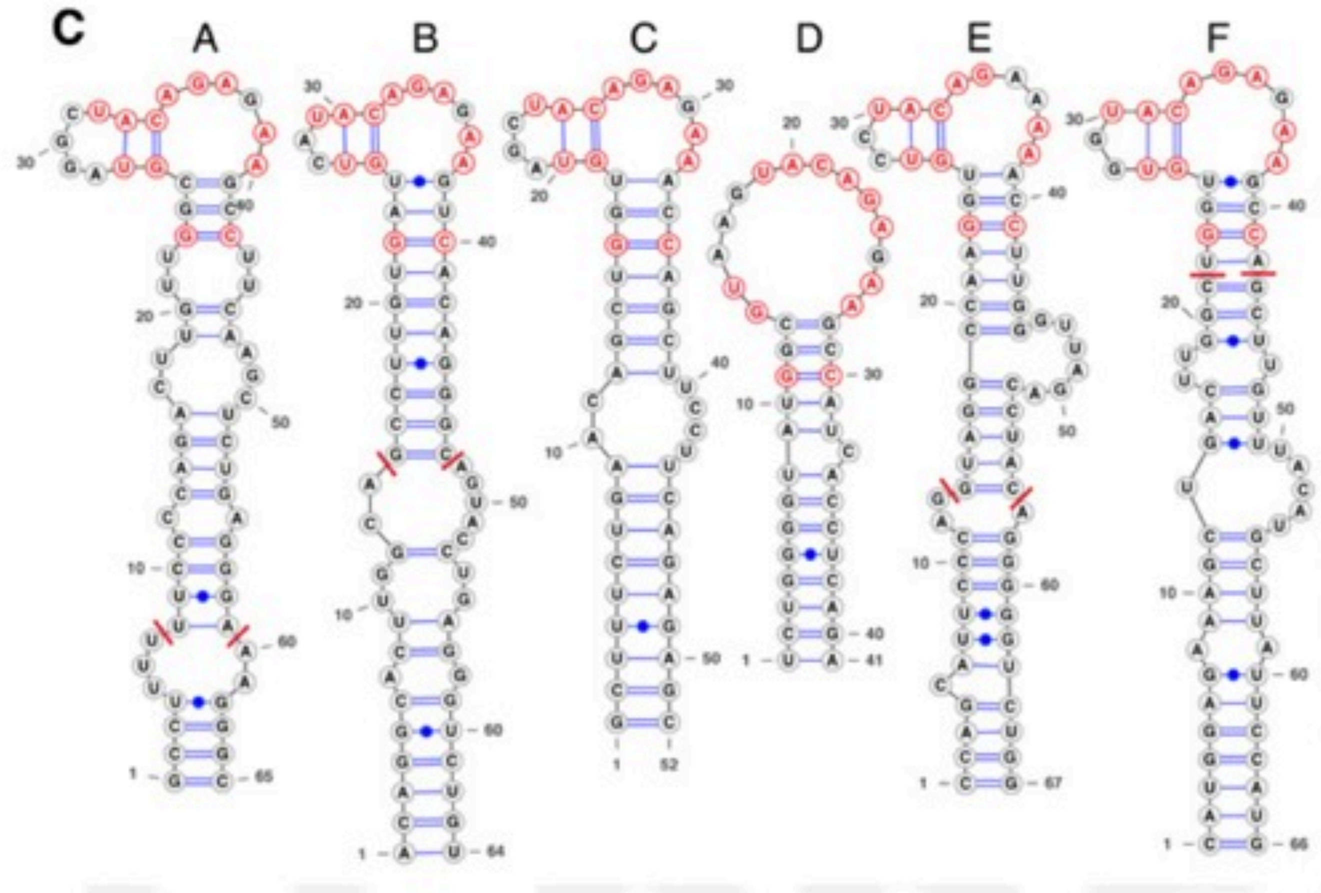
Loop motif shared between human members and down through vertebrates



```
                   42    1   3                   3    2    24
Human          D   UCUGGGGUAUGGCGUAAGUACAGAGAAGCCAUCACCUCAGA
Human          A   UCCCAGACUUGGCGUAGGUACAGAGAAGCCAAGCUCUGAGA
Human          B   ...GCCUUGUGAUGUCA-UACAGAGAAGUCAC-AGGGC...
Human          C   UCUGAAAGCUGGUGUAGCUACAGAGAAACCAGCUUUUCAGA
Human          E   ....GGCCAAGGUGUCC-UACAGAAAAACCUUGGGUU....
Human          F   .........UGGUGUG-GUACAGAGAAGCCA..........
                             *  **   ****** **  *


                   abcdefghijklm              mlkji hgfedcba
                   (((((((((((((..............))))).))))))))
Human              UCUGGGGUAUGGCGUAAGUACAGAGAAGCCAUCACCUCAGA
Guinea Pig         UCUGGGGUAUGGCGUAAGUACAGAGAAGCCAUCGCCUCAGA
Squirrel           UCUGAGGUAUGGUGUAAGUACAGAGAAGCCAUCACCUCAGA
Rabbit             UCUGGGGGAUGGCGUAAGUACAGAGAAGCCAUCUCCUCAGA
Hedgehog           UCUGAGGUAUGGCGUAAGUACAGAGAAGCCAUCACCUCAGA
Tenrec             U-GGGGUAUGGCUUAAGUACAGAGAAGCCCUCACCUCAGA
Sloth              UCUGGGGUAUGGUGUAAGUACAGAGAAGCCGUCACCUCAGA
Opossum            UCUGGGGUGUGGCGUGAGUACAGAGAAGCUAUCACCUCAGA
Lizard             U-UGGGACCGGGUGUGAGUACAGAGAAGCCCUUGUCUCAAA
X. tropicalis      UCUAGGCUUGGGCGUAAGUACAGAGUAGCCUUUGCCUU---
Tetraodon          UCUGAGGCCCGGCGUGGAUACAGAGAAGUCGGGCUUUCAGG
Fugu               UCUGAGGCCCGGCGUGGAUACAGAGAAGUCGGGCUGUCAGG
Stickleback        UCUGAGACGCAGCGUGGAUACAGAGAAGCUGUGGUUUCAGA
Medaka             UCUGGAACUCGGCGUGGAUACAGAGAAGCCGAUGUUUCAGA
Zebrafish          CUUGAGCCUUGGCGUCGGUACAGAAAAGCCGGGAUCUCAAG
                                 *   ****** **
```

Human structures with conserved motif



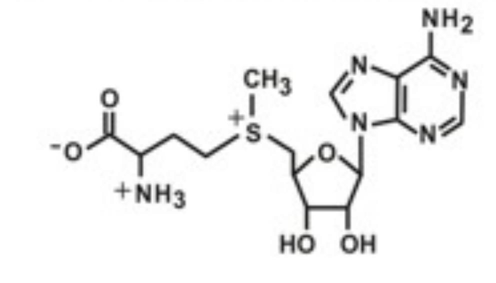
Brian J. Parker, et al. New families of human regulatory RNA structures identified by comparative analysis of vertebrate genomes. Genome Research (2011).

Post-transcriptional regulation of MAT2A

MAT2A: methionine adenosyltransferase II, alpha
MAT catalyzes the synthesis of SAM (adoMet)

Half-life of MAT2A transcript depends on SAM concetration
(Martínez-Chantar et al. J Biol Chem (2003))

SAM



Post-transcriptional regulation of MAT2A

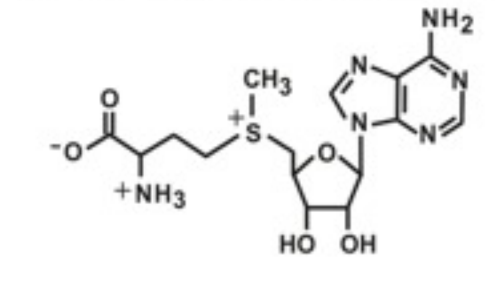
MAT2A: methionine adenosyltransferase II, alpha

MAT catalyzes the synthesis of SAM (adoMet)

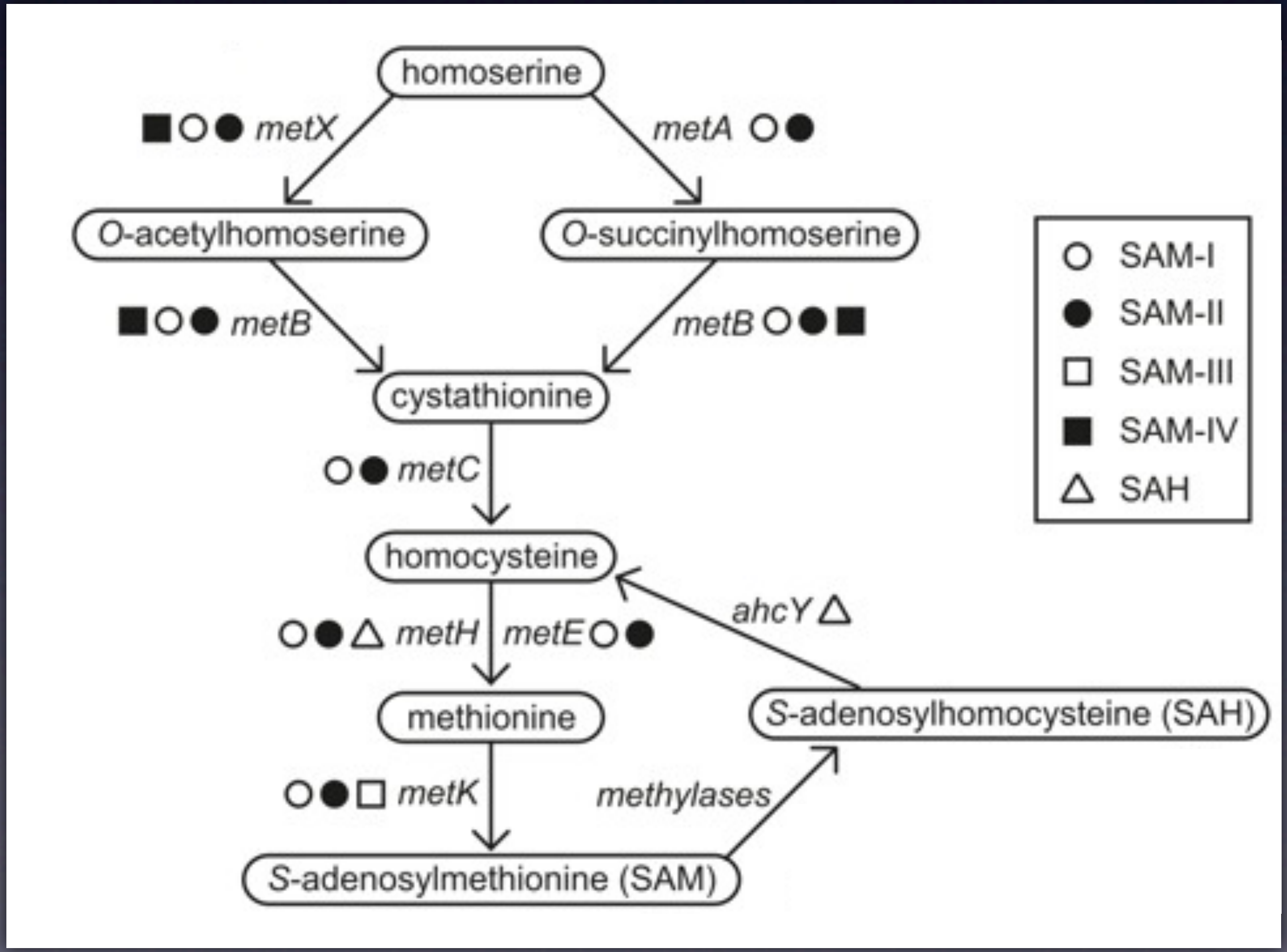
Half-life of MAT2A transcript depends on SAM concetration

(Martínez-Chantar et al. J Biol Chem (2003))

SAM

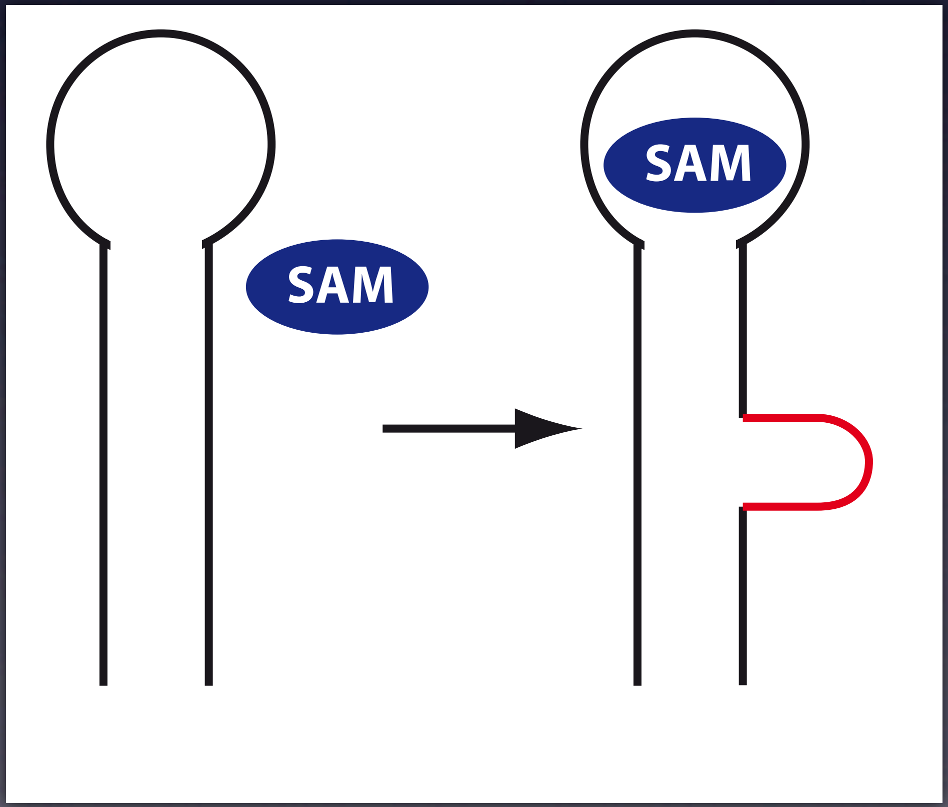


SAM riboswitches in bacteria



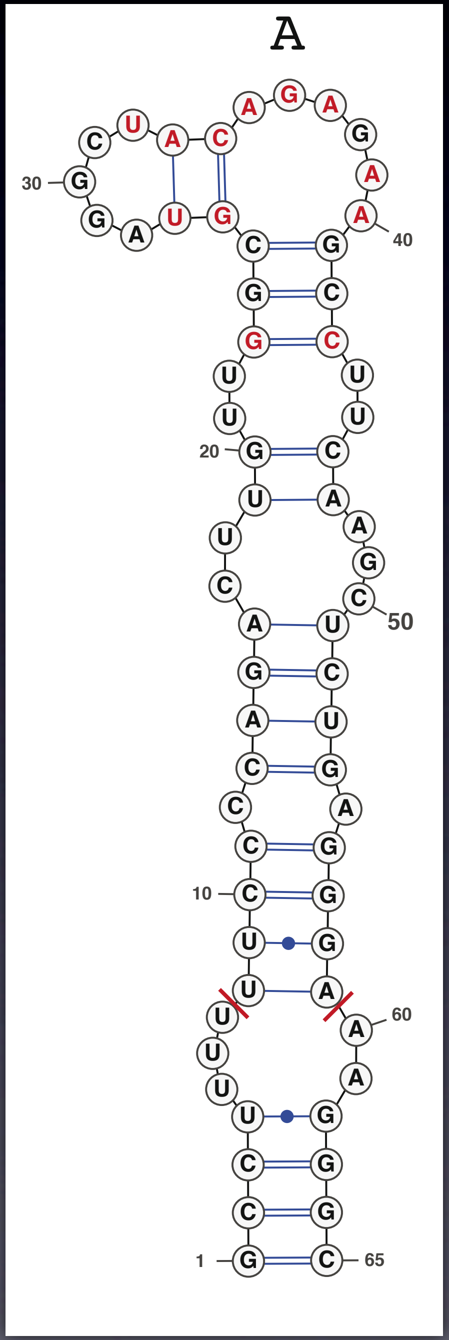
Wang and Breaker. Biochem Cell Biol (2008)

Riboswitches

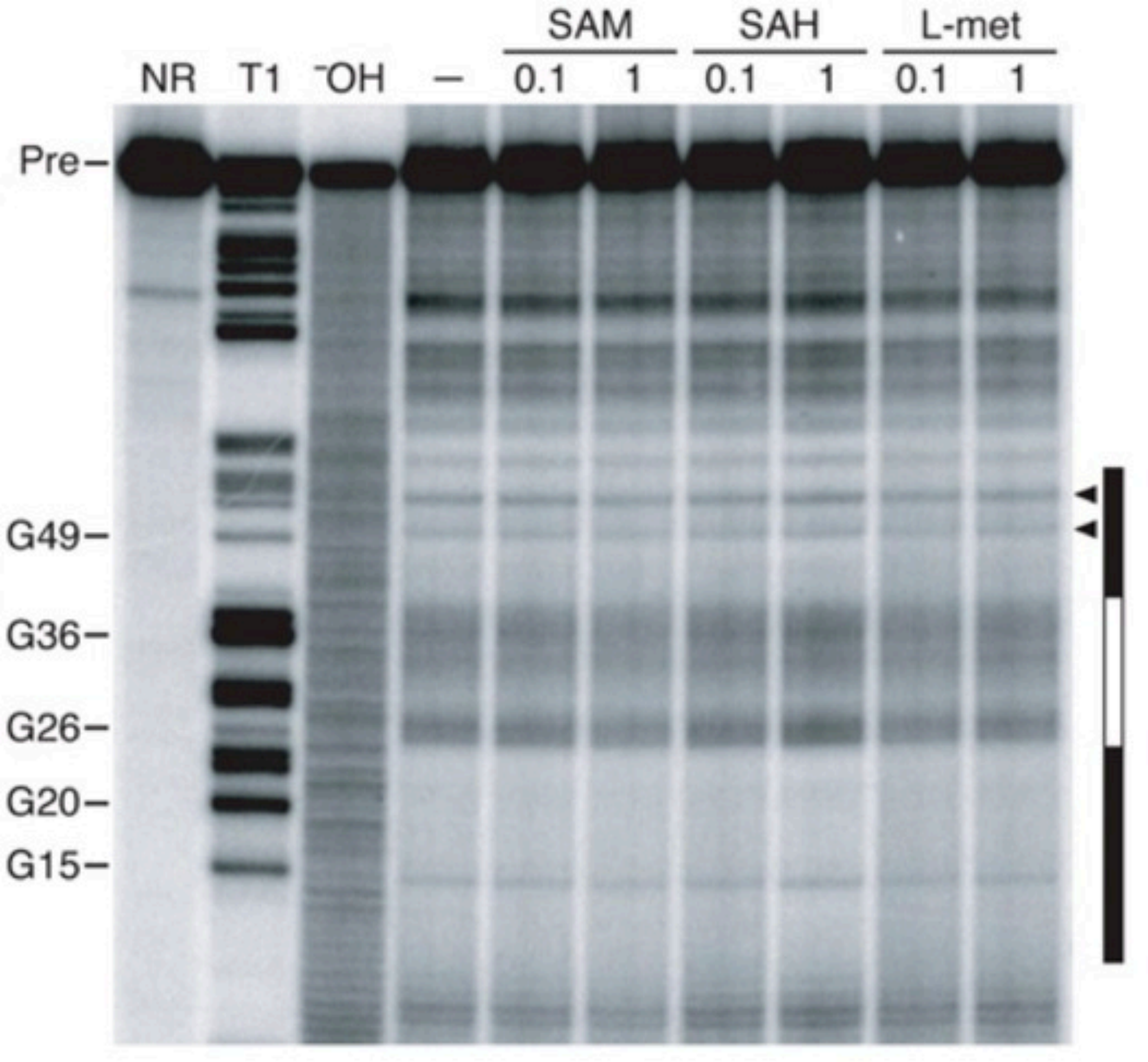


Human riboswitches?

Hairpin A



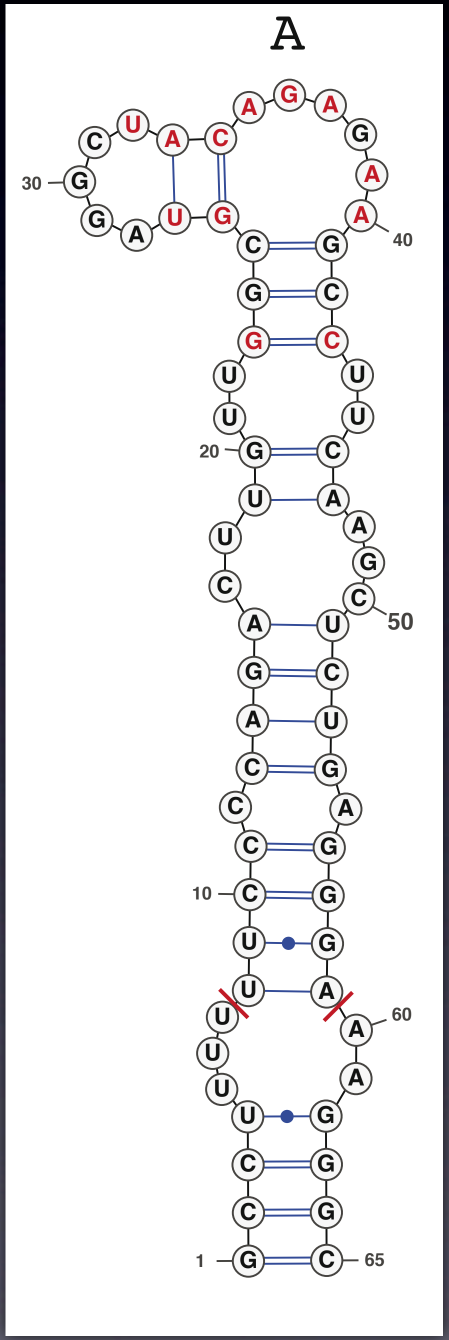
No structure change shown by in-line probing



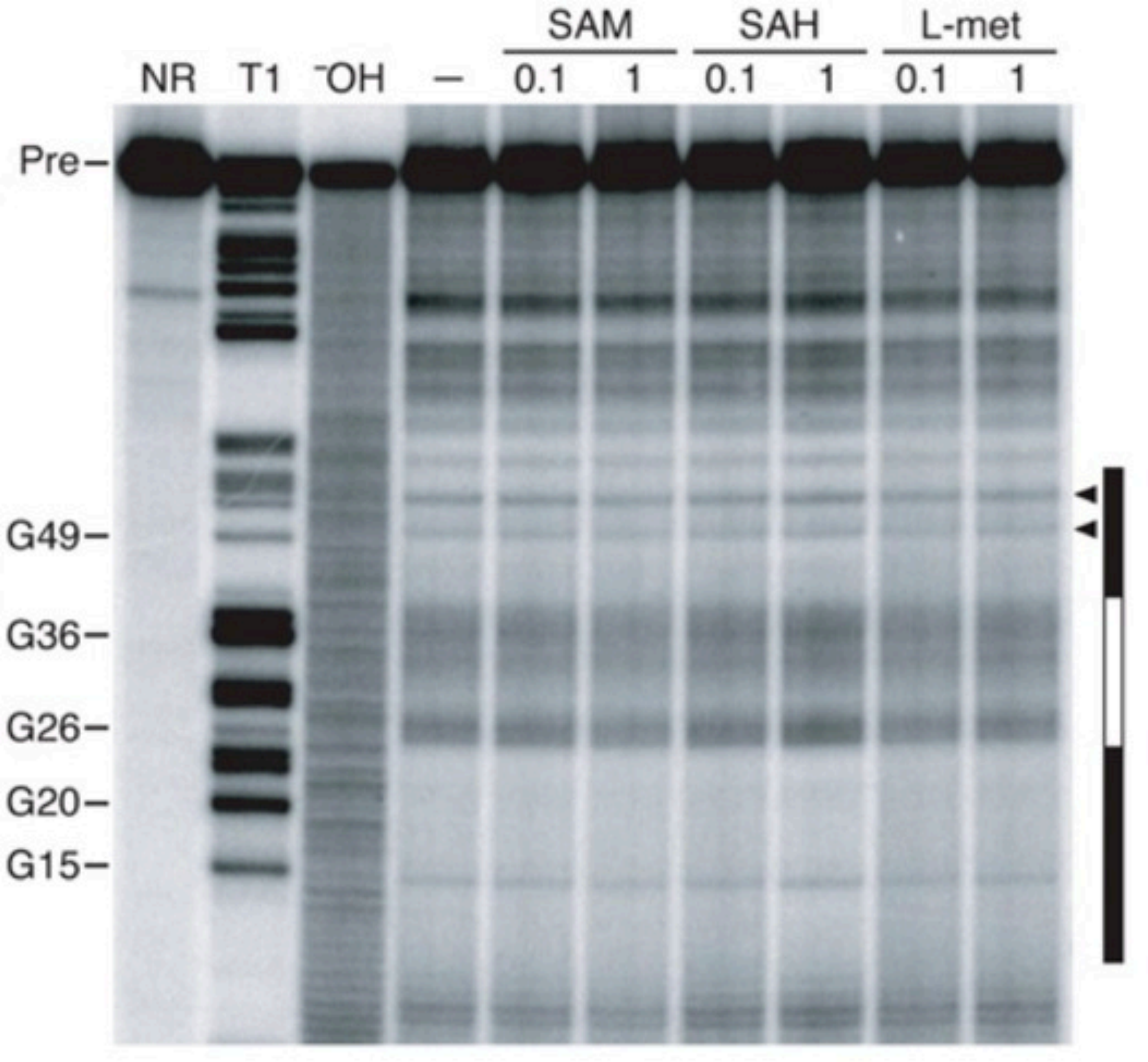
Experiments done by Adam Roth & Ronald Breaker (Yale).

Human riboswitches? Apparently not...

Hairpin A



No structure change shown by in-line probing

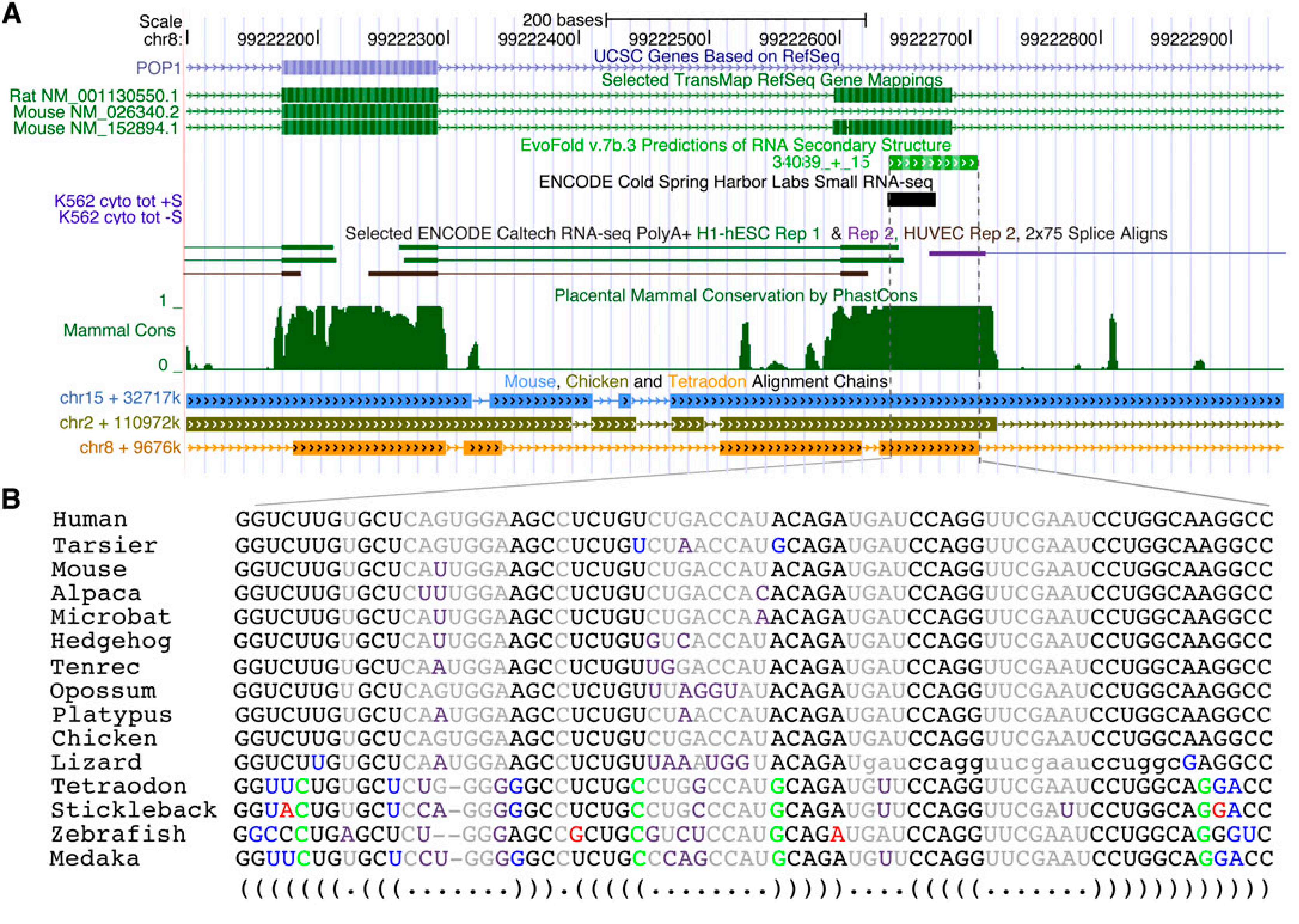


Experiments done by Adam Roth & Ronald Breaker (Yale).

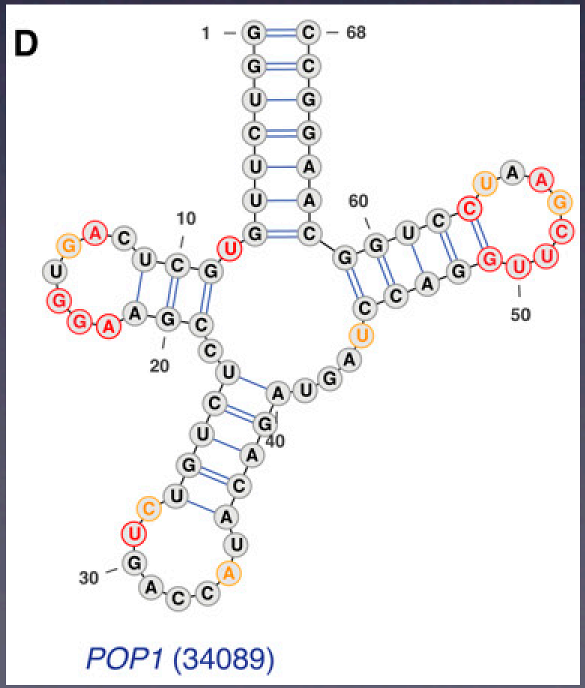
Example of auto-regulation?

POP1 is a ribonuclease, which is part of RNaseP that processes tRNAs.

tRNA-like structure in POP1 intron


A
Scale
chr8:
200 bases
99222200
99222300
99222400
99222500
99222600
99222700
99222800
99222900
POP1
UCSC Genes Based on RefSeq
Selected TransMap RefSeq Gene Mappings
Rat NM_001130550.1
Mouse NM_026340.2
Mouse NM_152894.1
EvoFold v.7b.3 Predictions of RNA Secondary Structure
34089_+_15
ENCODE Cold Spring Harbor Labs Small RNA-seq
K562 cyto tot +S
K562 cyto tot -S
Selected ENCODE Caltech RNA-seq PolyA+ H1-hESC Rep 1 & Rep 2, HUVEC Rep 2, 2x75 Splice Aligns
Placental Mammal Conservation by PhastCons
1
Mammal Cons
0
Mouse, Chicken and Tetraodon Alignment Chains
chr15 + 32717k
chr2 + 110972k
chr8 + 9676k
B

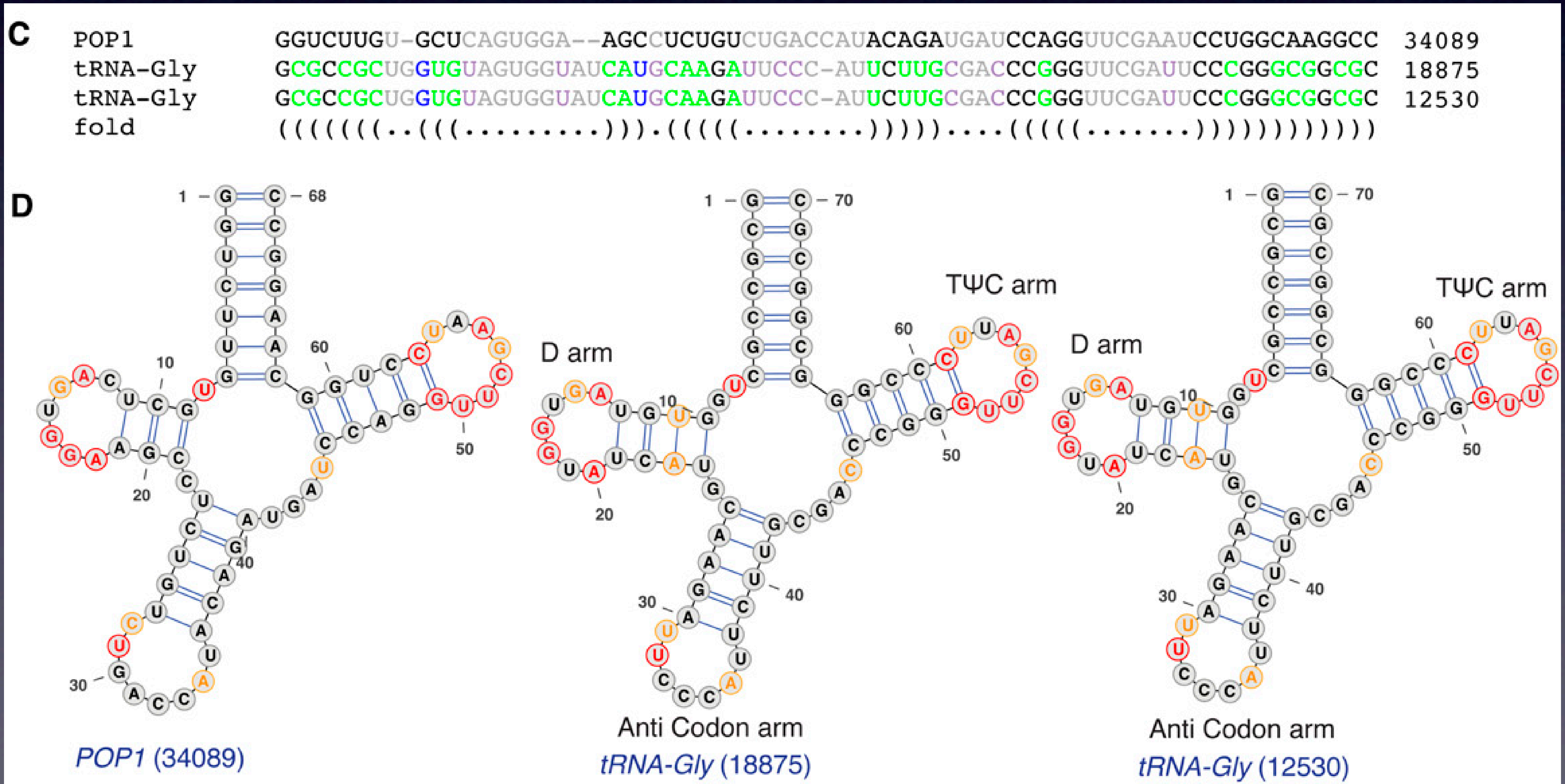
```
Human       GGUCUUGUGCUCAGUGGAAGCCUCUGUCUGACCAUACAGAUGAUCCAGGUUCGAAUCCUGGCAAGGCC
Tarsier     GGUCUUGUGCUCAGUGGAAGCCUCUGUCUAACCAUGCAGAUGAUCCAGGUUCGAAUCCUGGCAAGGCC
Mouse       GGUCUUGUGCUCAUUGGAAGCCUCUGUCUGACCAUACAGAUGAUCCAGGUUCGAAUCCUGGCAAGGCC
Alpaca      GGUCUUGUGCUCUUUGGAAGCCUCUGUCUGACCACACAGAUGAUCCAGGUUCGAAUCCUGGCAAGGCC
Microbat    GGUCUUGUGCUCAUUGGAAGCCUCUGUCUGACCAAACAGAUGAUCCAGGUUCGAAUCCUGGCAAGGCC
Hedgehog    GGUCUUGUGCUCAUUGGAAGCCUCUGUGUCACCAUACAGAUGAUCCAGGUUCGAAUCCUGGCAAGGCC
Tenrec      GGUCUUGUGCUCAAUGGAAGCCUCUGUUGGACCAUACAGAUGAUCCAGGUUCGAAUCCUGGCAAGGCC
Opossum     GGUCUUGU GCUCAGUGGAAGCCUCUGUUUAGGUAUACAGAUGAUCCAGGUUCGAAUCCUGGCAAGGCC
Platypus    GGUCUUGUGCUCAAUGGAAGCCUCUGUCUAACCAUACAGAUGAUCCAGGUUCGAAUCCUGGCAAGGCC
Chicken     GGUCUUGUGCUCAGUGGAAGCCUCUGUCUGACCAUACAGAUGAUCCAGGUUCGAAUCCUGGCAAGGCC
Lizard      GGUCUUGUGCUCAAUGGAAGCCUCUGUUAAAUGGUACAGAUgauccagguucgaauccuggcGAGGCC
Tetraodon   GGUUCUGUGCUCUG-GGGGCCUCUGCCUGGCCAUGCAGAUGUUCCAGGUUCGAAUCCUGGCAGGACC
Stickleback GGUACUGUGCUCCA-GGGGCCUCUGCCUGCCCAUGCAGAUGUUCCAGGUUCGAUUCCUGGCAGGACC
Zebrafish   GGCCCUGAGCUCU--GGGAGCCGCUGCGUCUCCAUGCAGAUGAUCCAGGUUCGAAUCCUGGCAGGGUC
Medaka      GGUUCUGUGCUCCU-GGGGGCCUCUGCCCAGCCAUGCAGAUGUUCCAGGUUCGAAUCCUGGCAGGACC
            (((((((.(((.......))).(((((........)))))....(((((.......))))))))))))
```



Parker et al. New families of human regulatory RNA structures identified by comparative analysis of vertebrate genomes. *in revision*.

Struture resembles tRNAs

POP1 structure groups together with tRNAs

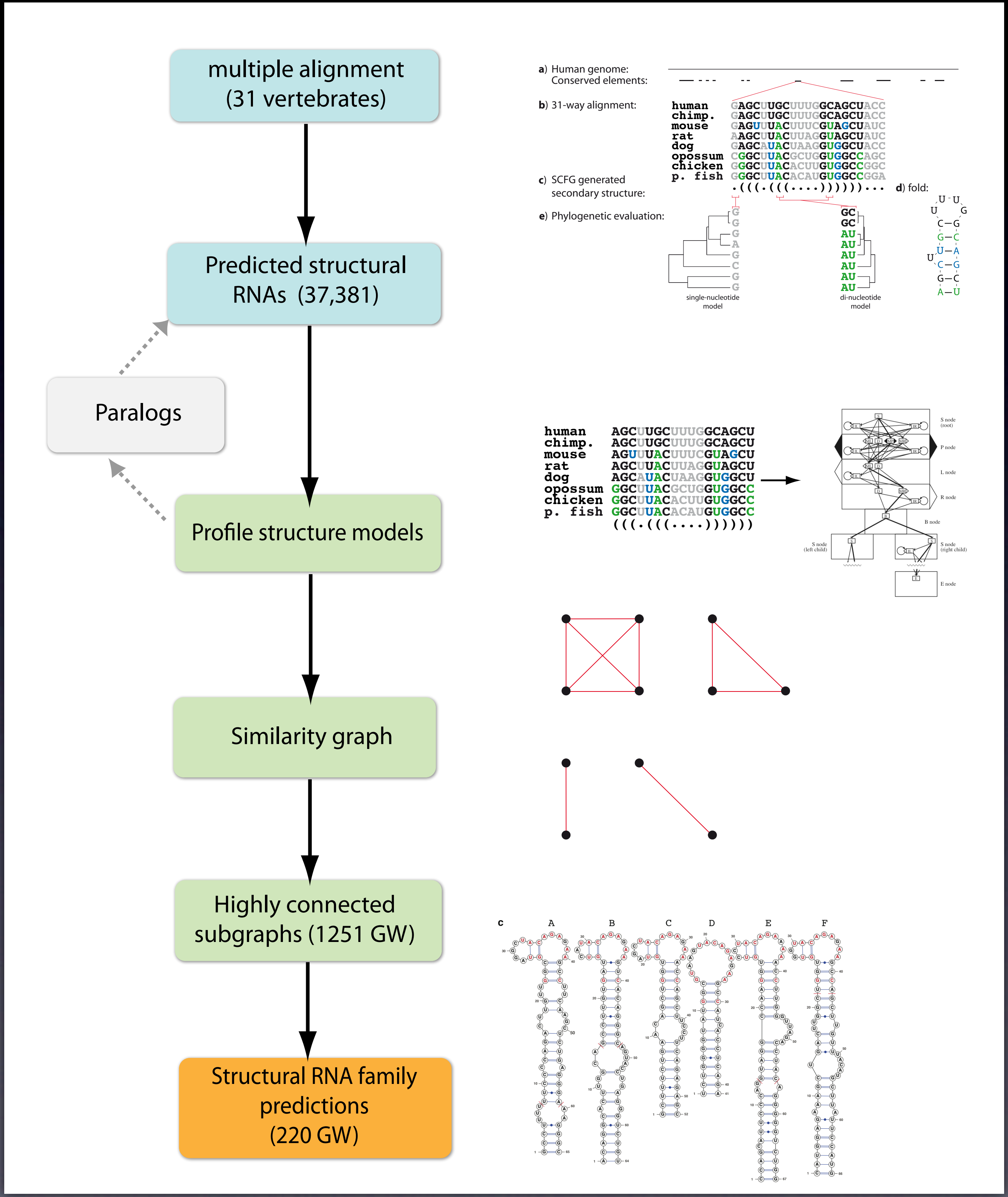


C

```
POP1       GGUCUUGU-GCUCAGUGGA--AGCCUCUGUCUGACCAUACAGAUGAUCCAGGUUCGAAUCCUGGCAAGGCC  34089
tRNA-Gly   GCGCCGCUGGUGUAGUGGUAUCAUGCAAGAUUCCC-AUUCUUGCGACCCGGGUUCGAUUCCCGGGCGGCGC  18875
tRNA-Gly   GCGCCGCUGGUGUAGUGGUAUCAUGCAAGAUUCCC-AUUCUUGCGACCCGGGUUCGAUUCCCGGGCGGCGC  12530
fold       (((((((..(((.........))).(((((........)))))....(((((.......))))))))))))
```

Parker et al. New families of human regulatory RNA structures identified by comparative analysis of vertebrate genomes. *in revision*.

EvoFam pipeline overview



multiple alignment (31 vertebrates)

Predicted structural RNAs (37,381)

Paralogs

Profile structure models

Similarity graph

Highly connected subgraphs (1251 GW)

Structural RNA family predictions (220 GW)

Future directions

- Make extensive use of deep genomic alignments (10K vertebrates project, etc)
- Exploit structure genome-wide structure probing data
- Integrate with expression data in cancer genomics settings
- Integrate with experimental evidence of binding sites of RNA binding proteins (HITS-CLIP, etc)

Acknowledgements

Structure families

Brian Parker (University of Copenhagen)
Ida Moltke (University of Copenhagen)
Jiayu Wen (University of Copenhagen)
Adam Roth (Yale)
Ronald Breaker (Yale)
Stefan Washietl (Broad)
Manolis Kellis (Broad)
Jakob Skou Pedersen (Aarhus University)

Holger Danske (left) & Brian Parker (right)

29 Mammals Sequencing and Analysis Consortium

Kerstin Lindblad-Toh
Manuel Garber
Or Zuk
Michael F. Lin
Brian J. Parker
Stefan Washietl
Pouya Kheradpour
Jason Ernst
Gregory Jordan
Evan Mauceli
Lucas D. Ward
Craig B. Lowe
Alisha K. Holloway
Michele Clamp
Sante Gnerre
Jessica Alfoldi
Kathryn Beal
Jean Chang
Hiram Clawson
James Cuff
Federica Di Palma
Stephen Fitzgerald
Paul Flicek

Mitchell Guttman
Melissa J. Hubisz
David B. Jaffe
Irwin Jungreis
W. James Kent
Dennis Kostka
Marcia Lara
Andre L. Martins
Tim Massingham
Ida Moltke
Brian J. Raney
Matthew D. Rasmussen
Jim Robinson
Alexander Stark
Albert J. Vilella
Jiayu Wen
Xiaohui Xie
Michael C. Zody
Broad Institute Sequencing Platform and Whole Genome Assembly Team

Kim C. Worley
Christie L. Kovar
Donna M. Muzny
Richard A. Gibbs
Baylor College of Medicine Human Genome Sequencing Center
Sequencing Team
Wesley C. Warren
Elaine R. Mardis
George M. Weinstock
Richard K. Wilson
Genome Institute at Washington
University
Ewan Birney
Elliott H. Margulies
Javier Herrero
Eric D. Green
David Haussler
Adam Siepel
Nick Goldman
Katherine S. Pollard
Jakob S. Pedersen
Eric S. Lander
Manolis Kellis

Benasque 2009

Eric Westhof
Zasha Weinberg

Support

Novo Nordisk Foundation (BJP)
The Danish Council for Independent Research | Medical Sciences (JSP)
Lundbeckfonden (JSP)

Brian J. Parker et al. New families of human regulatory RNA structures identified by comparative analysis of vertebrate genomes. Genome Research (2011). http://moma.ki.au.dk/prj/mammals/
Kerstin Lindblad-Toh, Manuel Garber, Or Zuk, Michael F. Lin, Brian J. Parker, et al. A high-resolution map of human evolutionary constraint using 29 mammals. Nature (2011).