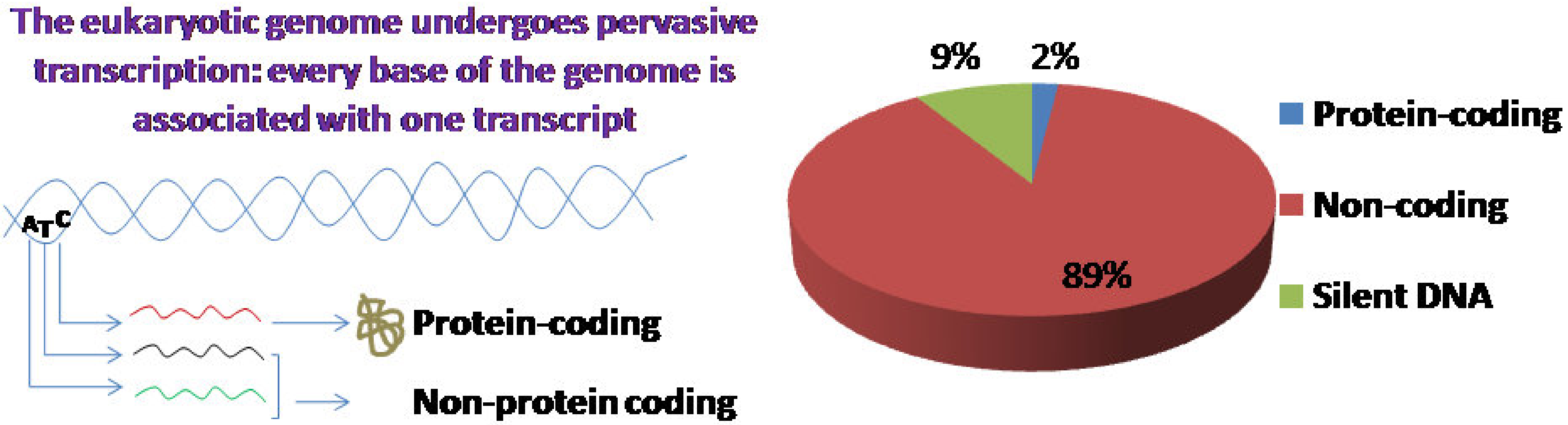
EECS730: Introduction to Bioinformatics

Lecture 11: Non-coding RNA discovery



The eukaryotic genome undergoes pervasive transcription: every base of the genome is associated with one transcript

http://desktop.jncasr.ac.in/uploaded/mrsrao/Slide1.jpg

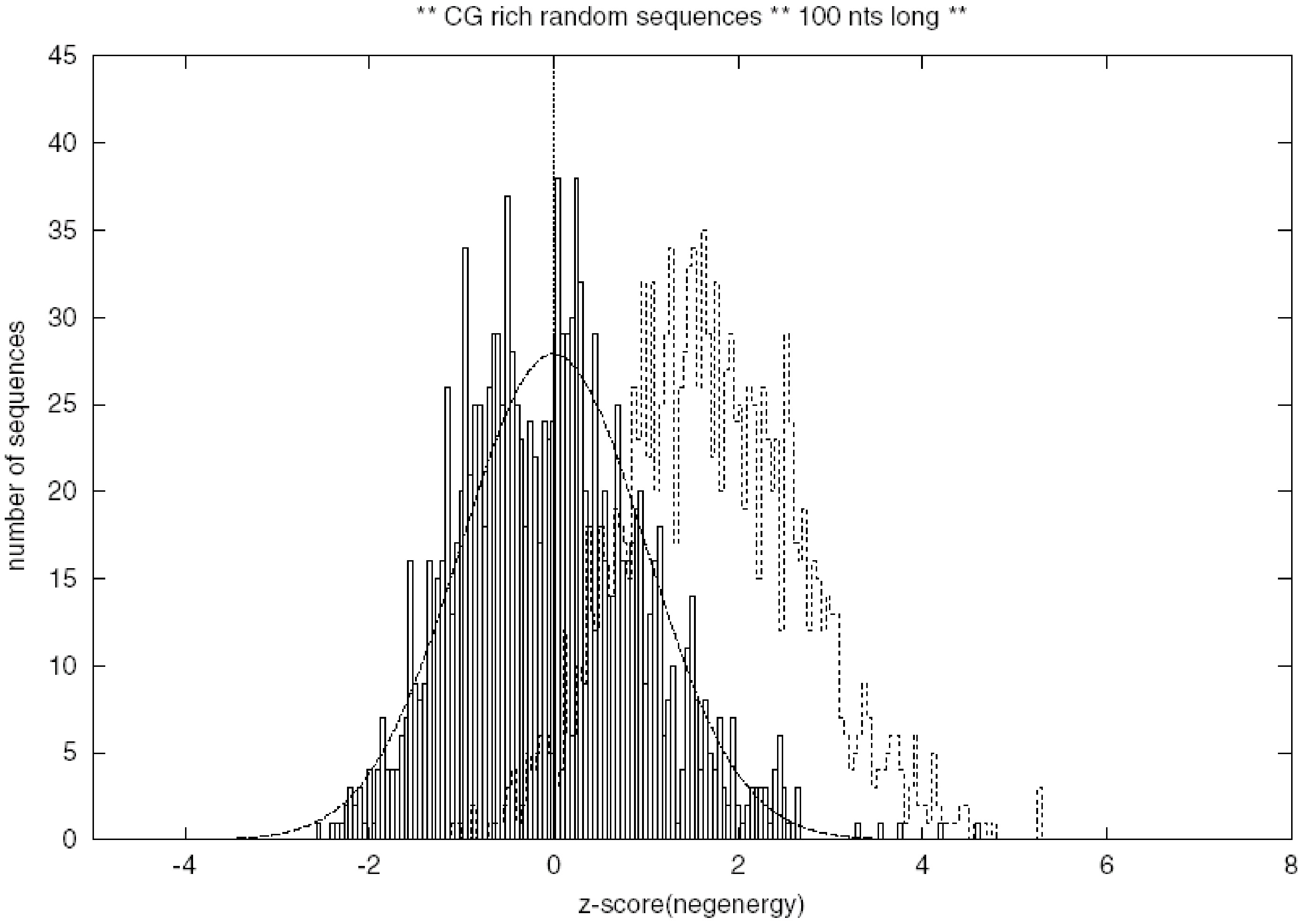
Slides adapted from Dr. Shaojie Zhang (University of Central Florida)

Problem: how can we predict noncoding RNA genes from the genome

- We know that we can do this for protein-coding genes (gene-finding)
- Using HMMs that summarize the gene features
- However, noncoding RNAs are in general harder to detect
- No codon preference information available

Stable secondary structure?

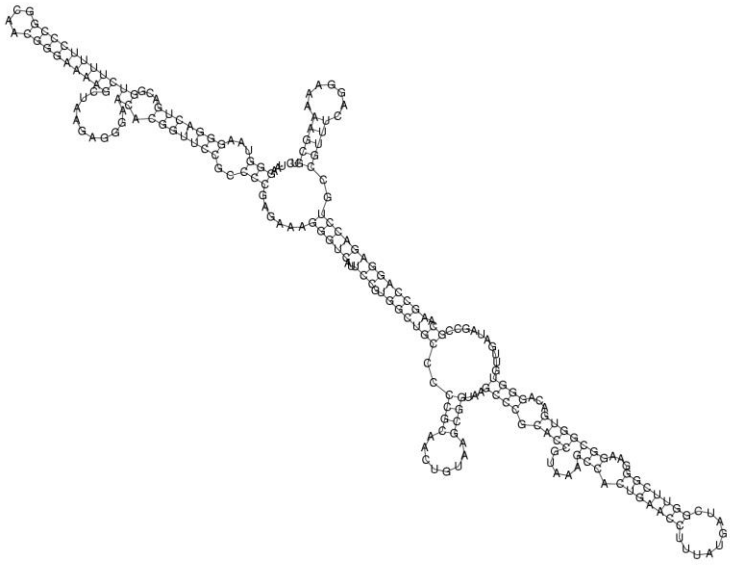
- The stability of ncRNA secondary structure is not sufficiently different from the predicted stability of a random sequence. [Rivas and Eddy *Bioinformatics* (2000)].

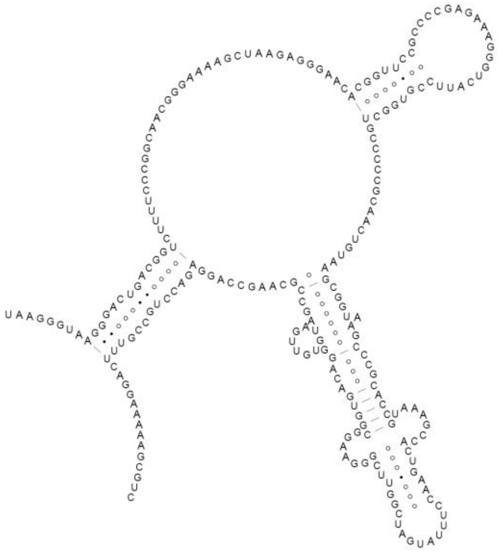


RNA folding

- Algorithms/programs to compute the minimum energy:
  - Nussinov et al (1978), Waterman (1978), Smith and Waterman (1978), and Zuker and Sankoff (1984).
  - **Mfold** (Zuker 2003) and **RNAfold (ViennaRNA)** (Hofacker 2003).

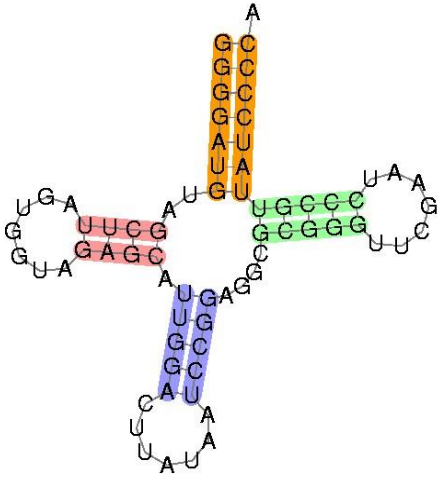
- RNA folding via energy minimization has its shortcomings:
  - Prediction depends on correct energy parameters.
  - Sometimes, the true structure does not have the minimum energy.

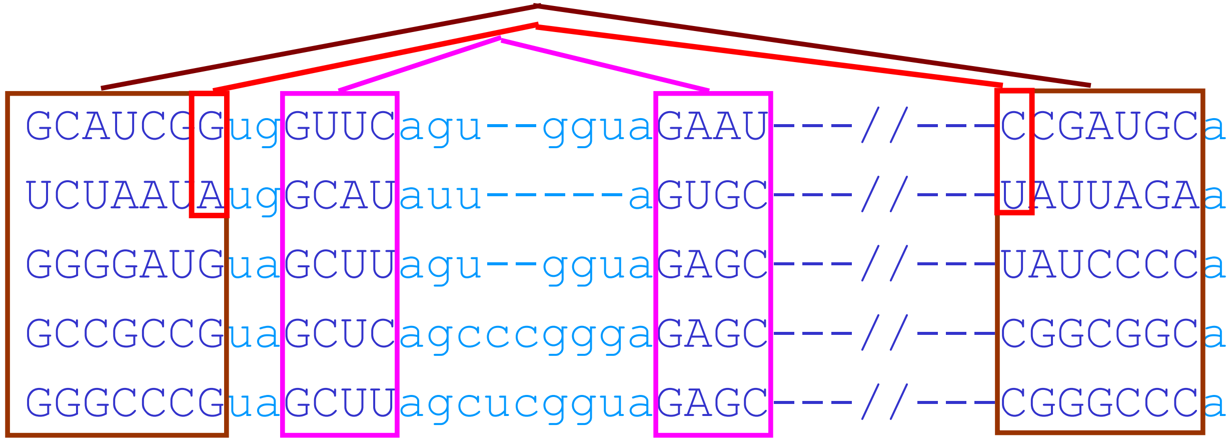




Other information to use?

- Covarying mutations found from the multiple sequence alignment is a strong indication of RNA secondary structure





```
GCAUCGGugGUUCagu--gguaGAAU---//---CCGAUGCa
UCUAAUAugGCAUauu-----aGUGC---//---UAUUAGAa
GGGGAUGuaGCUUagu--gguaGAGC---//---UAUCCCCa
GCCGCCGuaGCUCagcccgggaGAGC---//---CGGCGGCa
GGGCCCGuaGCUUagcucgguaGAGC---//---CGGGCCCa
```

Incorporating covarying mutation information

- If we have correct multiple alignments, looking for covarying mutations and finding consensus structure is a good way to do structure prediction.
    - **RNAalifold** (Hofacker et al. 2002)
    - The consensus structure prediction is more accurate.
    - To find energetically stable consensus structure is more statistically significant.
    - Still compute the MFE.
    - Covariance information is incorporated into the energy model by rewarding compensatory and consistent mutations.

$$E_{i,j} = \min \left\{ E_{i,j-1};\ \min_{\substack{k:\ i+m<k\leq j \\ \Pi_{ik}=1}} E_{i+1,k-1} + E_{k+1,j} + \beta_{ik} \right\}$$

Incorporating covarying mutation

- Take into account covariance contribution:

$$d_{ij}^{\alpha,\beta} = 2 - \delta\left(a_i^\alpha, a_i^\beta\right) - \delta\left(a_j^\alpha, a_j^\beta\right)$$

$$C_{ij} = \frac{1}{\binom{N}{2}} \sum_{\alpha<\beta} d_{ij}^{\alpha,\beta} \Pi_{ij}^\alpha \Pi_{ij}^\beta$$

- Take into account inconsistent sequences:

$$q_{ij} = 1 - \frac{1}{N} \sum_\alpha \left\{ \Pi_{ij}^\alpha + \delta(a_i^\alpha, \mathrm{gap})\delta\left(a_j^\alpha, \mathrm{gap}\right) \right\}$$

where $\delta(a', a'') = 1$, if $a' = a''$ and 0, otherwise.

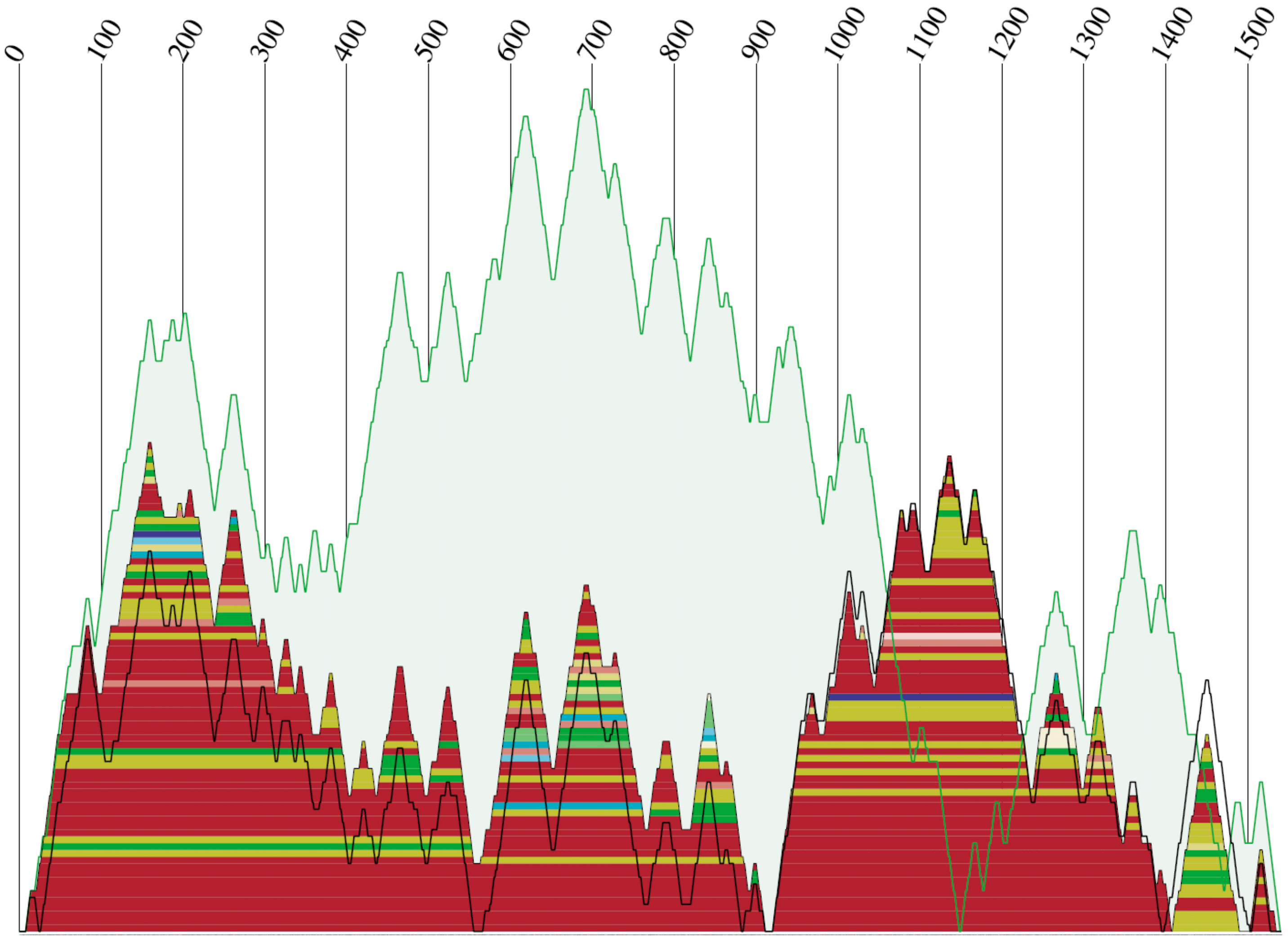
- Put together:

$$B_{ij} = C_{ij} - \phi_1 q_{ij}$$

$$\Pi_{ij} = \breve{1}$$

if sequence positions i and j can form a base-pair, i.e. if (x_i, x_j) is in the set of allowed base-pairs $B = \{GC, CG, AU, UA, GU, UG\}$, and $\Pi_{ii} = 0$ if x_i and x_i cannot pair.

Mountain plot of 16s rRNA

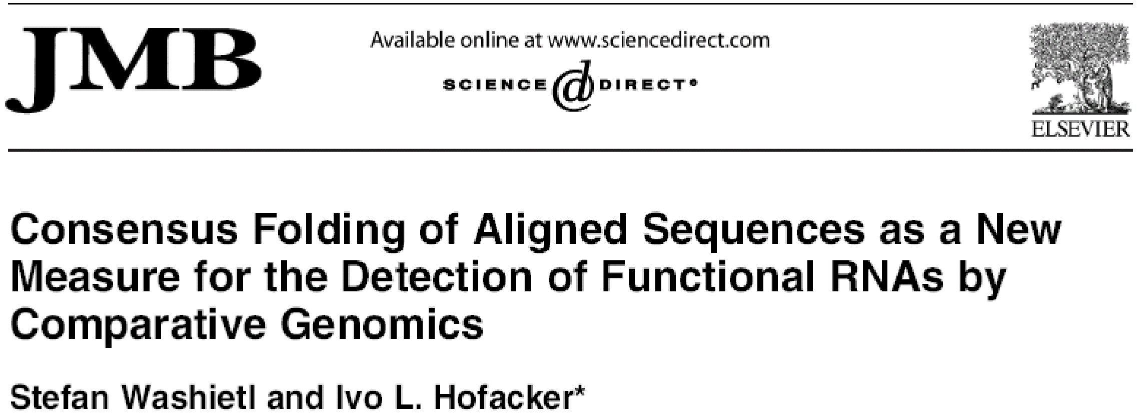


De novo detection of RNA elements

- To find energetically stable consensus structure is more statistically significant.

- MFE can be used to compute the statistical significance.
  - MFE: *m*
  - Mean: *ц*
  - Standard deviation: *δ*
  - Z-score: *z = (m- ц)/ δ*

- We need randomize the multiple sequence alignment
  - Shuffle the columns of the input alignment
    - Not destroy the gap structure.
    - Certain sequence pattern.



JMB

Available online at www.sciencedirect.com

SCIENCE @ DIRECT®

ELSEVIER

Consensus Folding of Aligned Sequences as a New Measure for the Detection of Functional RNAs by Comparative Genomics

Stefan Washietl and Ivo L. Hofacker*

AlifoldZ

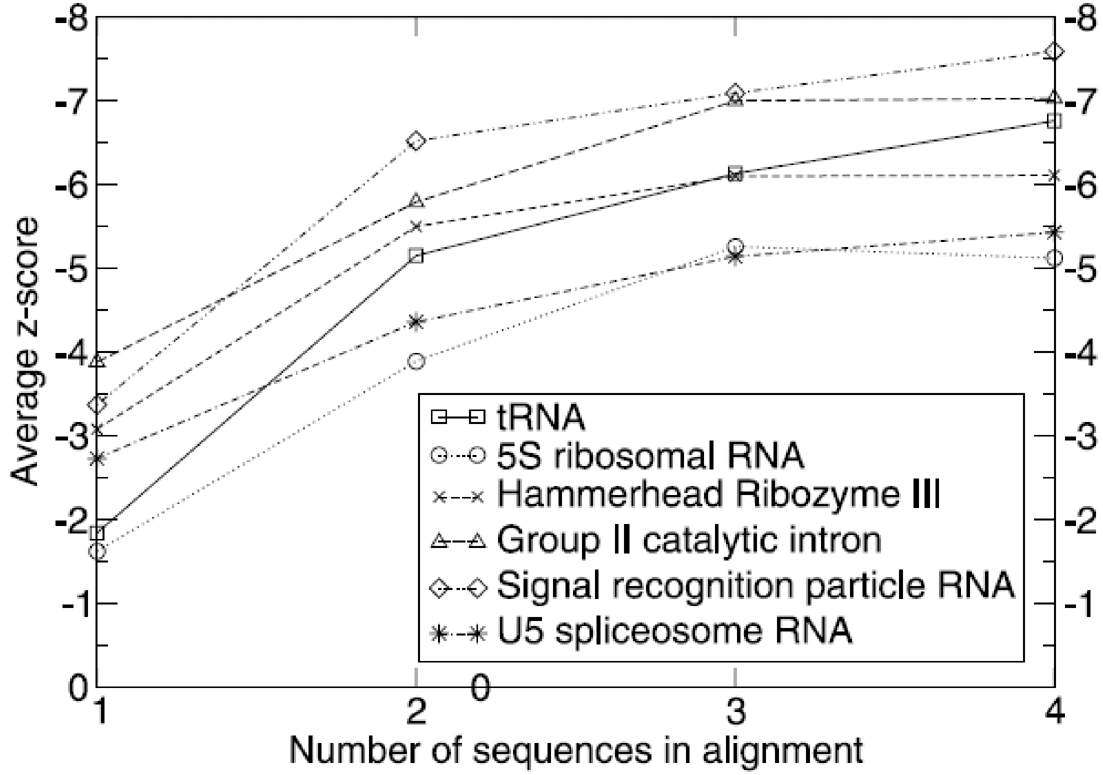
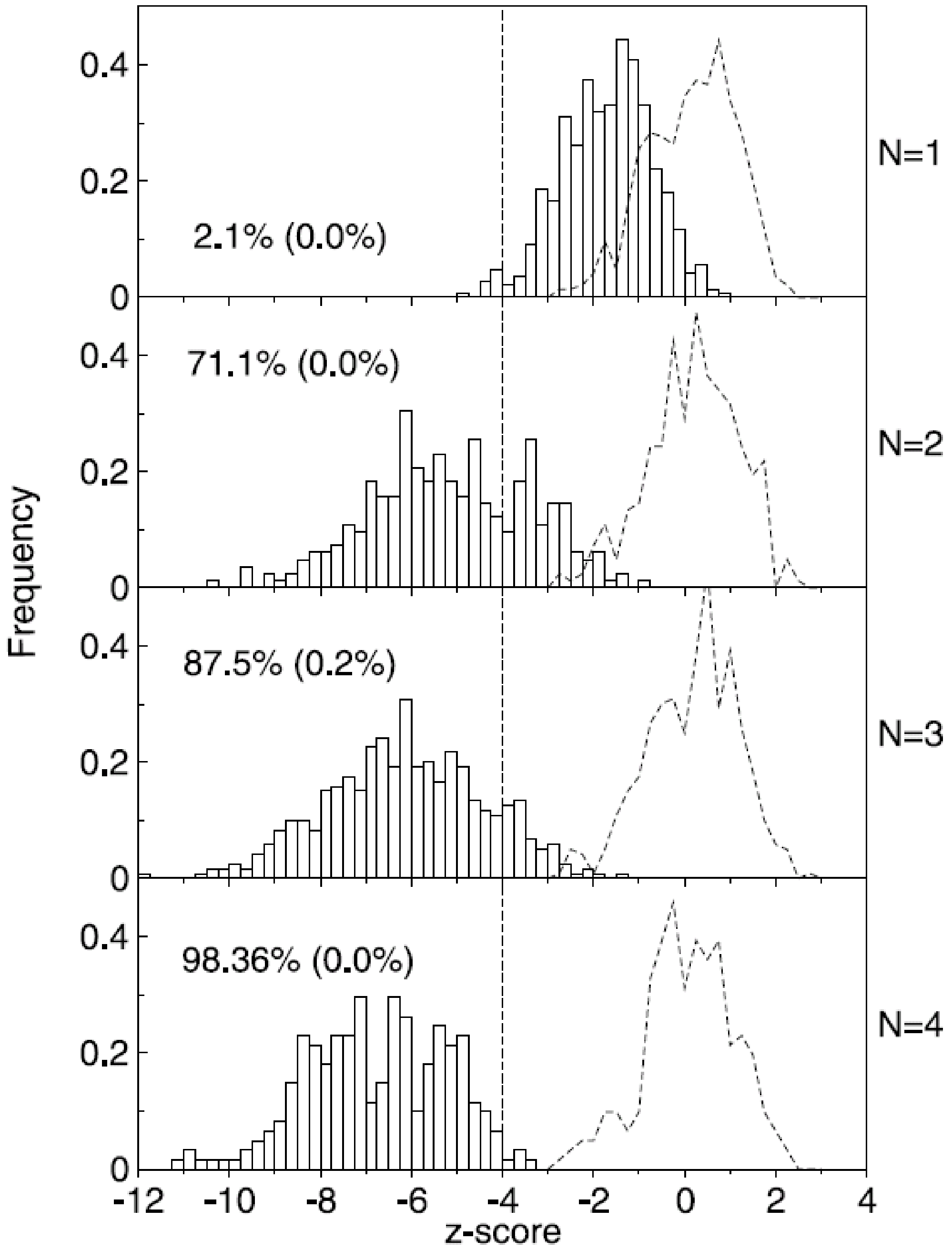
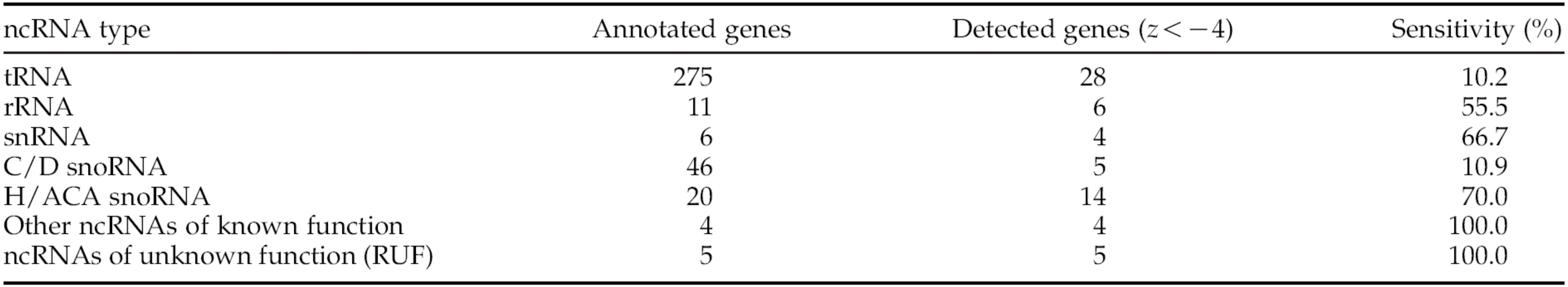


Figure 1. Mean *z*-scores of various RNA types dependent on the number of sequences in alignment. $N=1$ means RNAfold predictions for single sequences. Mean pairwise identities of the alignments are between 65% and 85%. See Table 1 for more details.



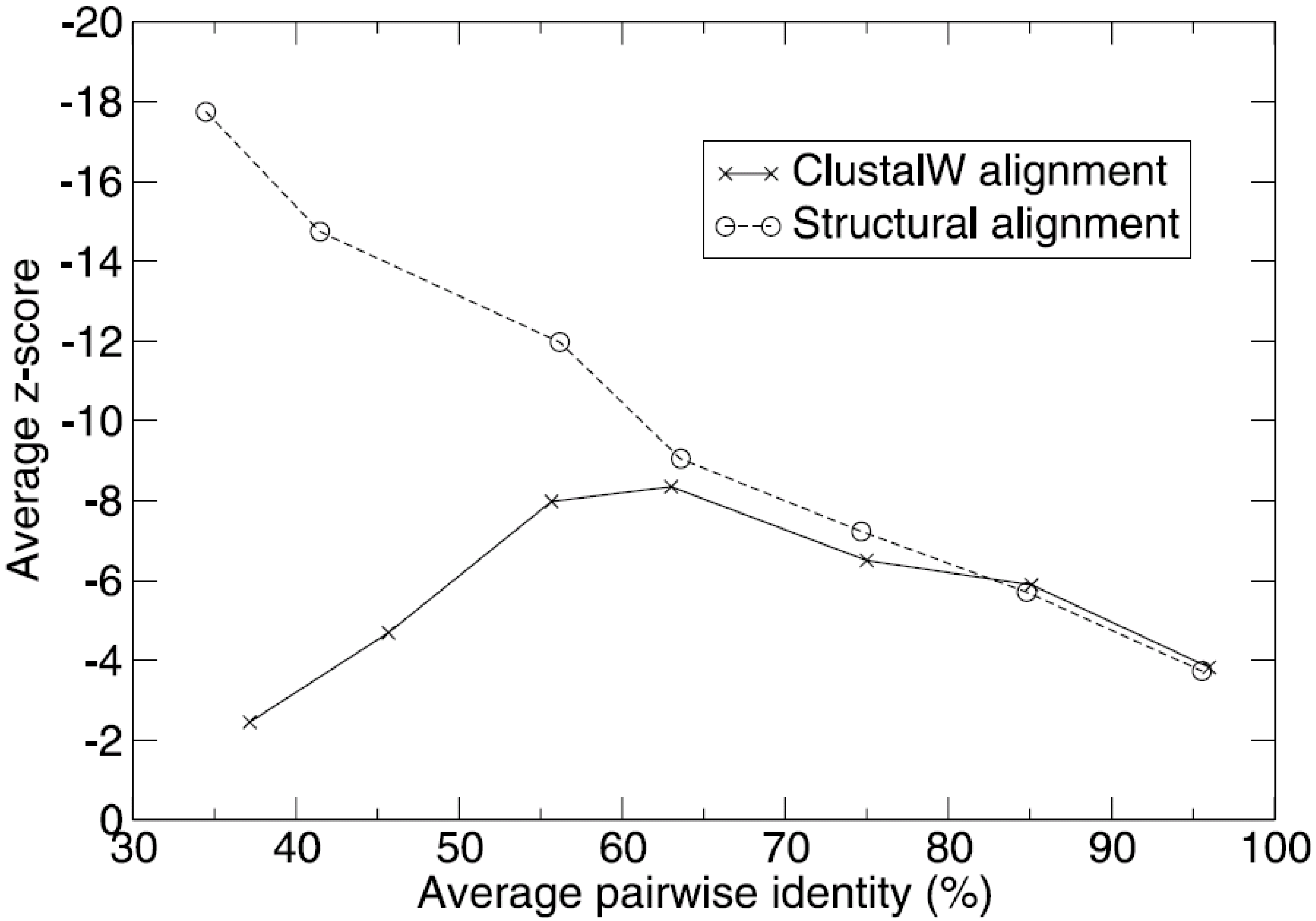
Real data performance

- Use MultiPipMaker to generate the multiple alignment of *S. cerevisiae* and other 6 related yeast genome.
- Extracted the regions of annotated ncRNAs
- Refine the poorly aligned regions
- Window size = 150, slide 20.
- False-positive rate: 0.25%.
- 30 CPU days.



ncRNA type	Annotated genes	Detected genes ($z < -4$)	Sensitivity (%)
tRNA	275	28	10.2
rRNA	11	6	55.5
snRNA	6	4	66.7
C/D snoRNA	46	5	10.9
H/ACA snoRNA	20	14	70.0
Other ncRNAs of known function	4	4	100.0
ncRNAs of unknown function (RUF)	5	5	100.0

Problem remains



(Based on pairwise alignments of SRP RNAs)

- We need good multiple alignments to correctly predict secondary structures.

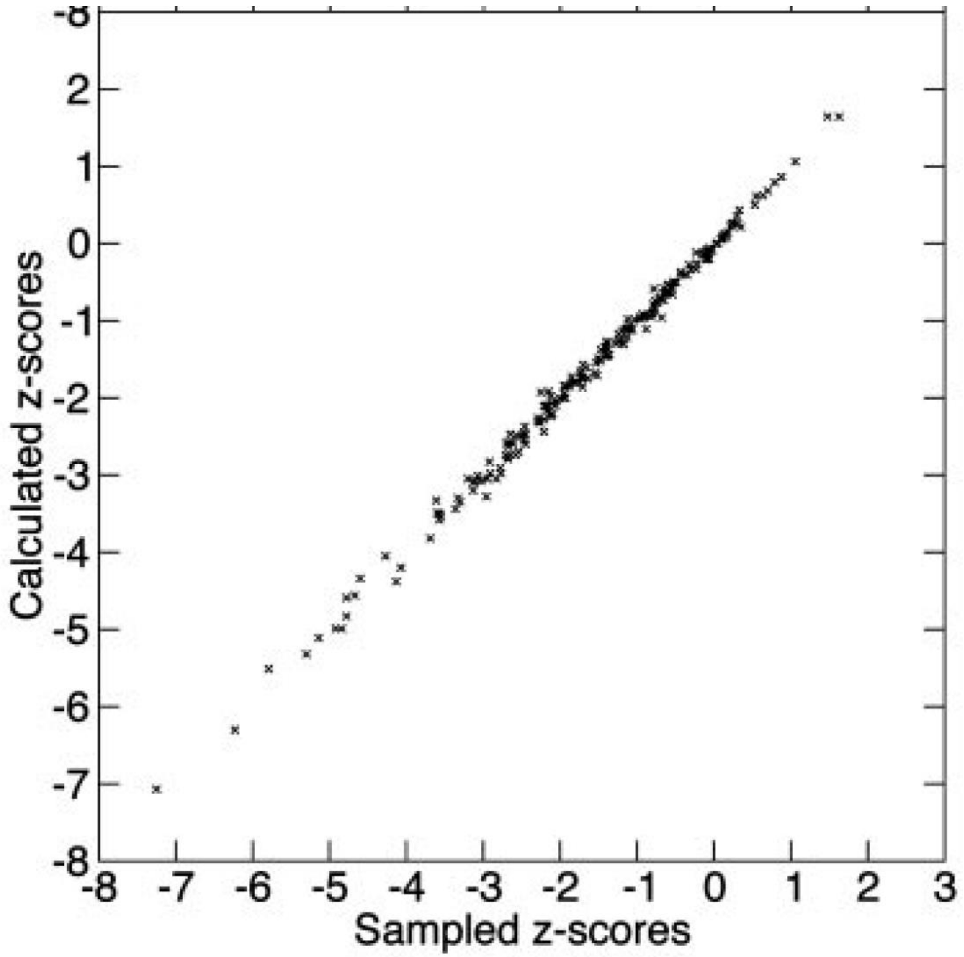
- We need to know the correct secondary structures to generate good multiple alignments.

- Solution:
    - Use Simultaneous Alignment and Folding (Sankoff Algorithm); computational intensive
    - Only apply on RNA sequences who have the "right" sequence similarity (between 60-95%)

RNAz (PNAS, 2005)

- z-score (for individual sequence)
  - Using Support Vector Machine (SVM) regression.
  - Using >10,000 point to define the independent variables (4-variables in total).
    - different length.
    - different base composition (GC/AT, A/T, G/C ratio).
  - Compute Mean (ų) and standard deviation (*δ)* for each data point to define the dependent variable
  - Compute the MFE of the sequence, and compute Z-score: *z = (m- ų)/ δ*
  - For an alignment, using the mean of the z-scores.

Z-score estimation

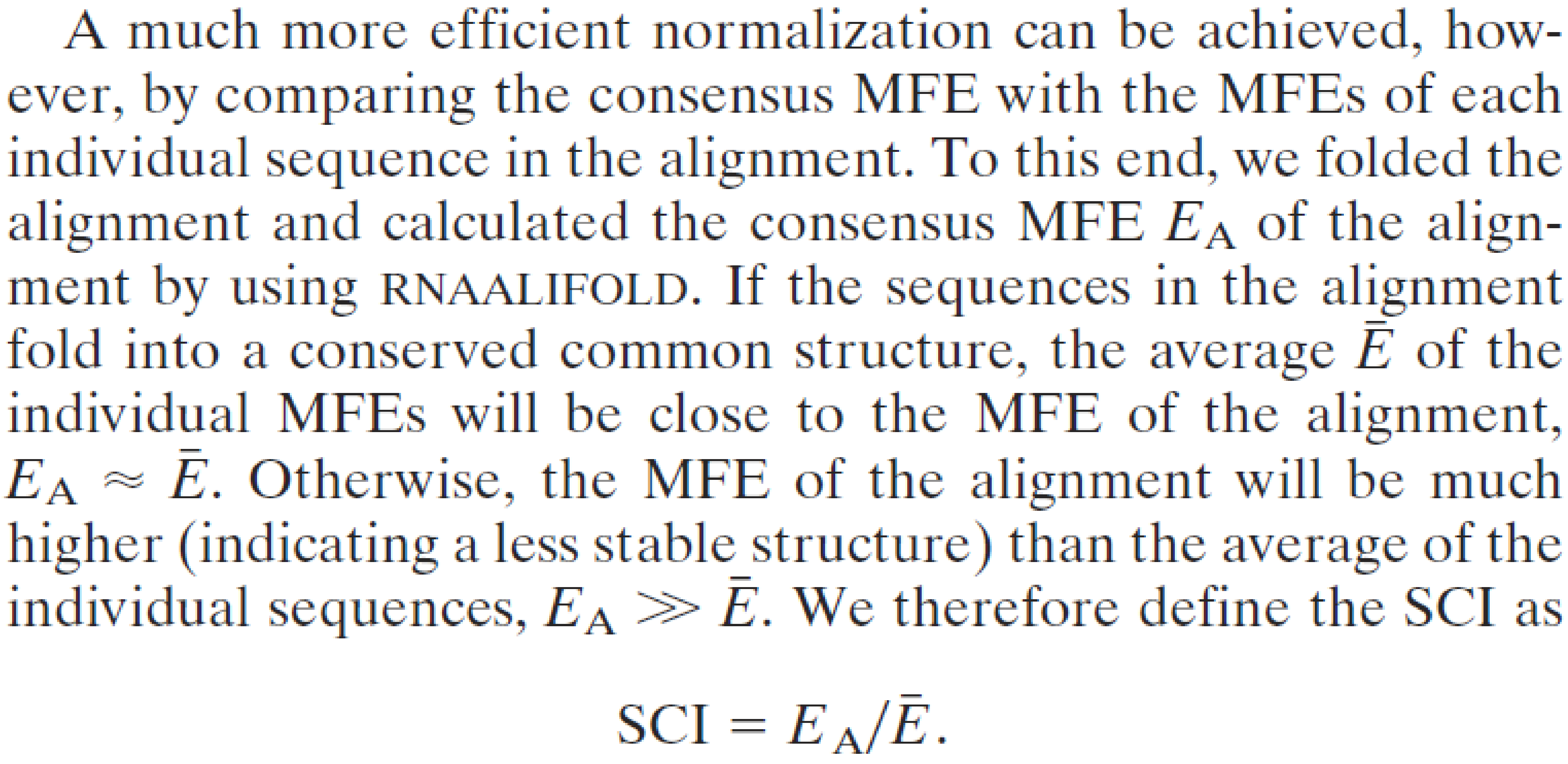


RNAz (classifying true/false noncoding RNA)

- Estimate a probability (*P*) if the alignment is classified as a structured RNA, based on
  - SCI
  - z-score
  - Average pairwise identity
  - Number of sequences.

- It is also done by SVM.

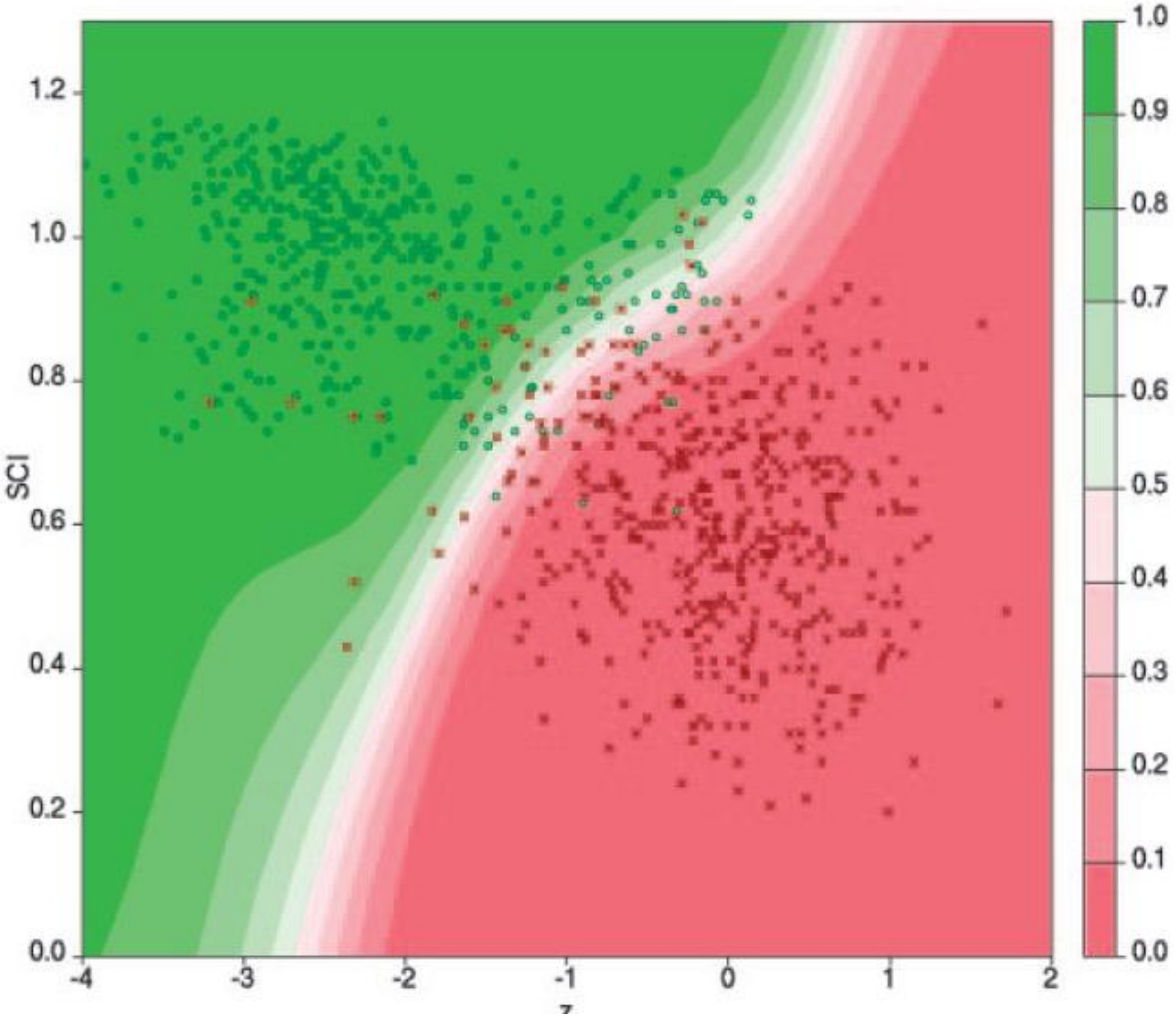
SCI (structure conservation index)

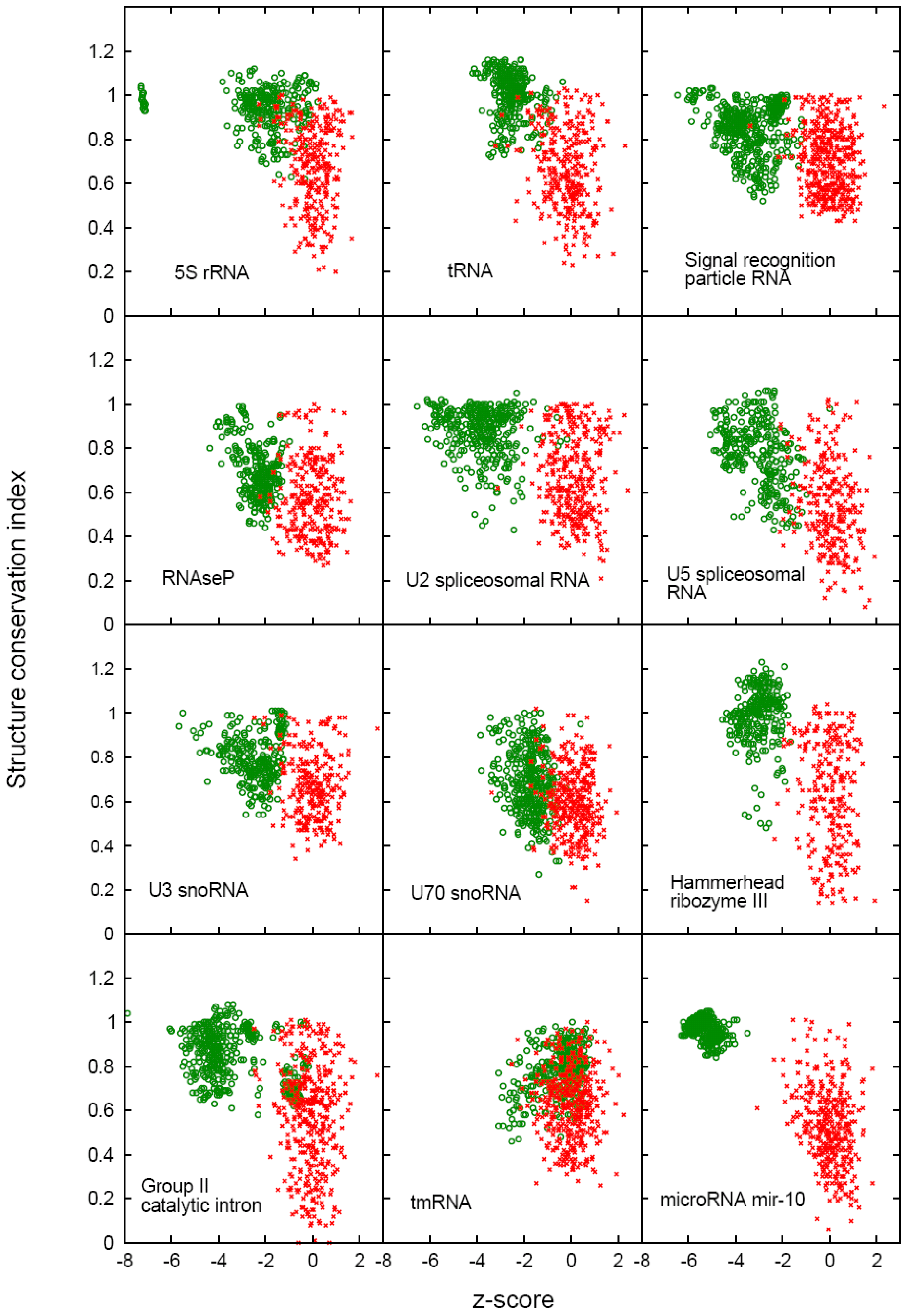


A much more efficient normalization can be achieved, however, by comparing the consensus MFE with the MFEs of each individual sequence in the alignment. To this end, we folded the alignment and calculated the consensus MFE E_{A} of the alignment by using RNAALIFOLD. If the sequences in the alignment fold into a conserved common structure, the average $\bar{E}$ of the individual MFEs will be close to the MFE of the alignment, $E_{\mathrm{A}} \approx \bar{E}$. Otherwise, the MFE of the alignment will be much higher (indicating a less stable structure) than the average of the individual sequences, $E_{\mathrm{A}} \gg \bar{E}$. We therefore define the SCI as

$$\mathrm{SCI} = E_{\mathrm{A}}/\bar{E}.$$

Classification based on z scores and SCI by a SVM




5S rRNA
tRNA
Signal recognition particle RNA
RNAseP
U2 spliceosomal RNA
U5 spliceosomal RNA
U3 snoRNA
U70 snoRNA
Hammerhead ribozyme III
Group II catalytic intron
tmRNA
microRNA mir-10
Structure conservation index
0
0.2
0.4
0.6
0.8
1
1.2
z-score
-8
-6
-4
-2
0
2

Performance of RNAz

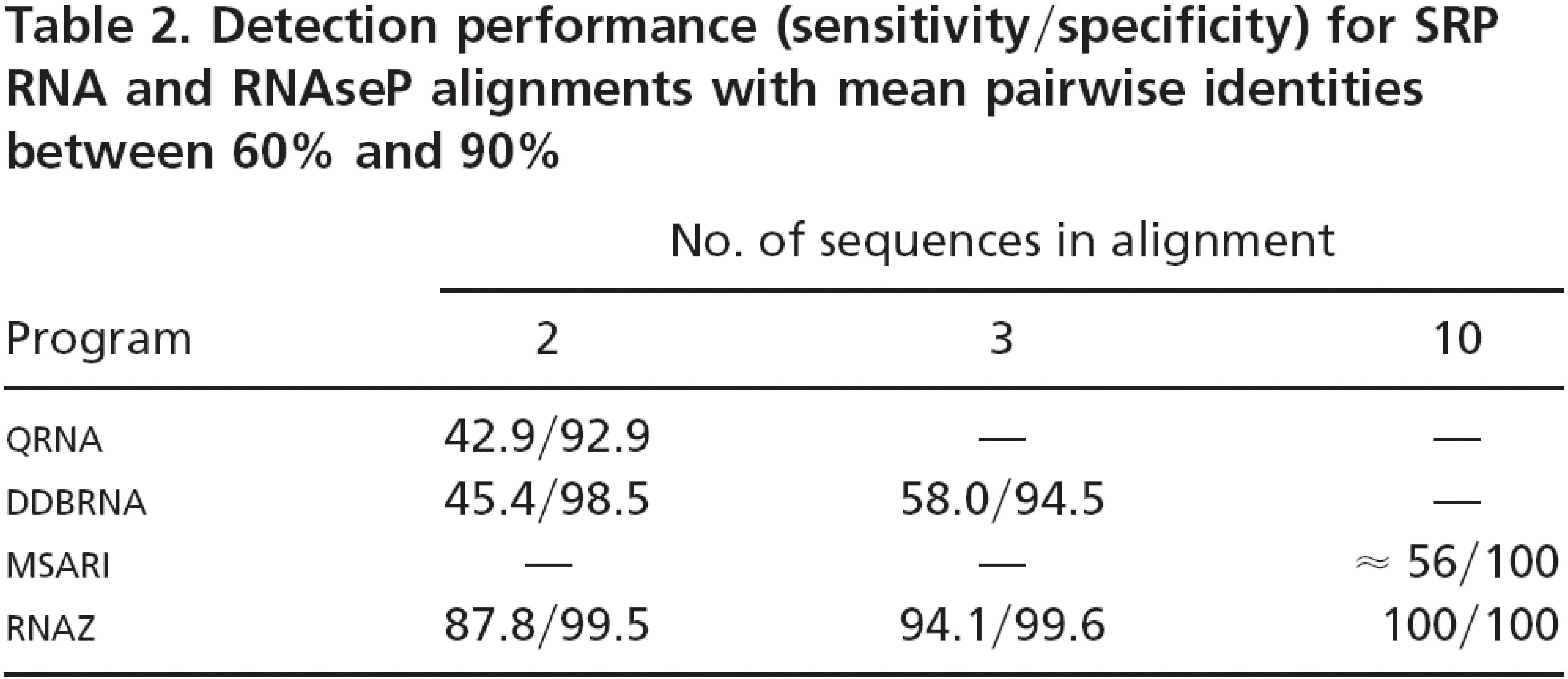


Table 2. Detection performance (sensitivity/specificity) for SRP RNA and RNAseP alignments with mean pairwise identities between 60% and 90%

	No. of sequences in alignment		
Program	2	3	10
QRNA	42.9/92.9	—	—
DDBRNA	45.4/98.5	58.0/94.5	—
MSARI	—	—	≈ 56/100
RNAZ	87.8/99.5	94.1/99.6	100/100

Using RNAz to scan the human genome

- *Nature Biotechnology* 23, 1383 - 1390 (Nov. 2005), “**Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome**”
- Input:
  - Genome-wide alignments of vertebrates from UCSC genome browser.
  - Using PhastCons program to find the most conserved
  - Adjacent conserved regions (<50 distances) are joined together.
  - All regions > 50 bps.
  - Remove all “known genes” and “Refseq genes”
- Output:
  - Predicted structured RNA elements in the human genomes using RNAz

Results

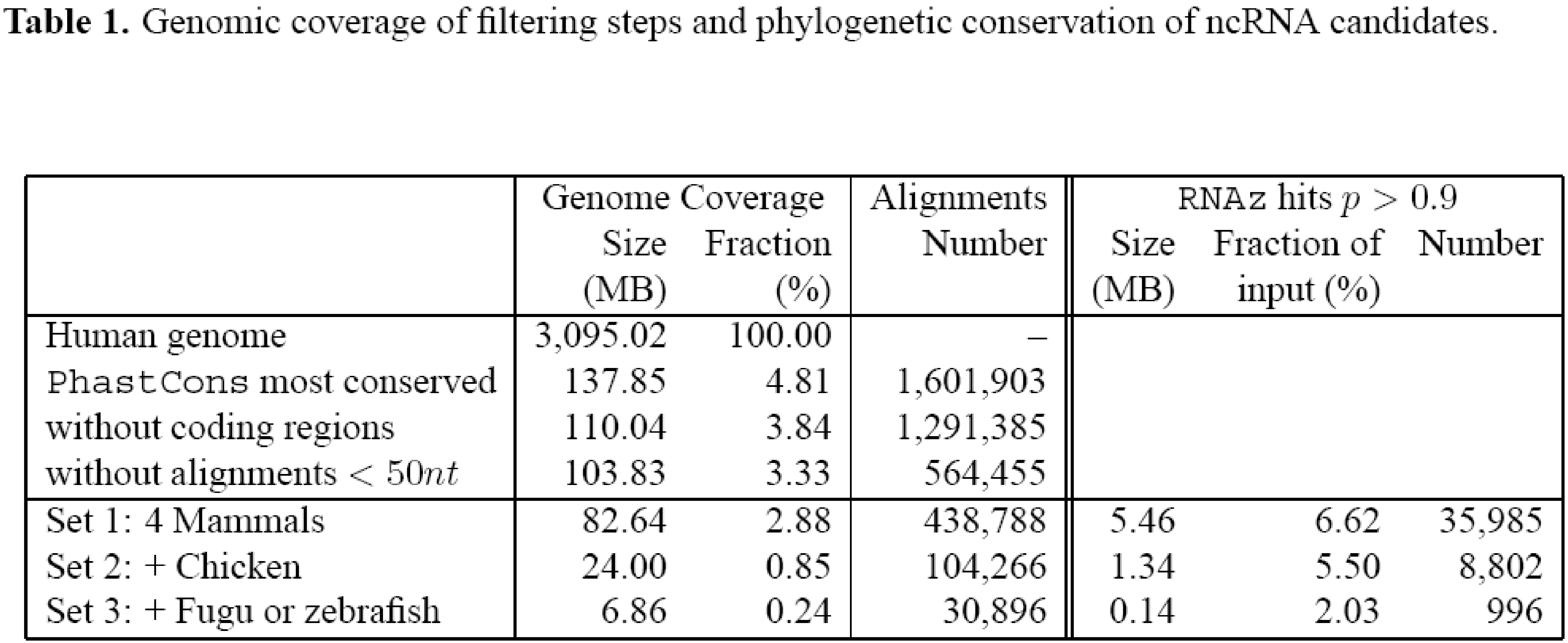
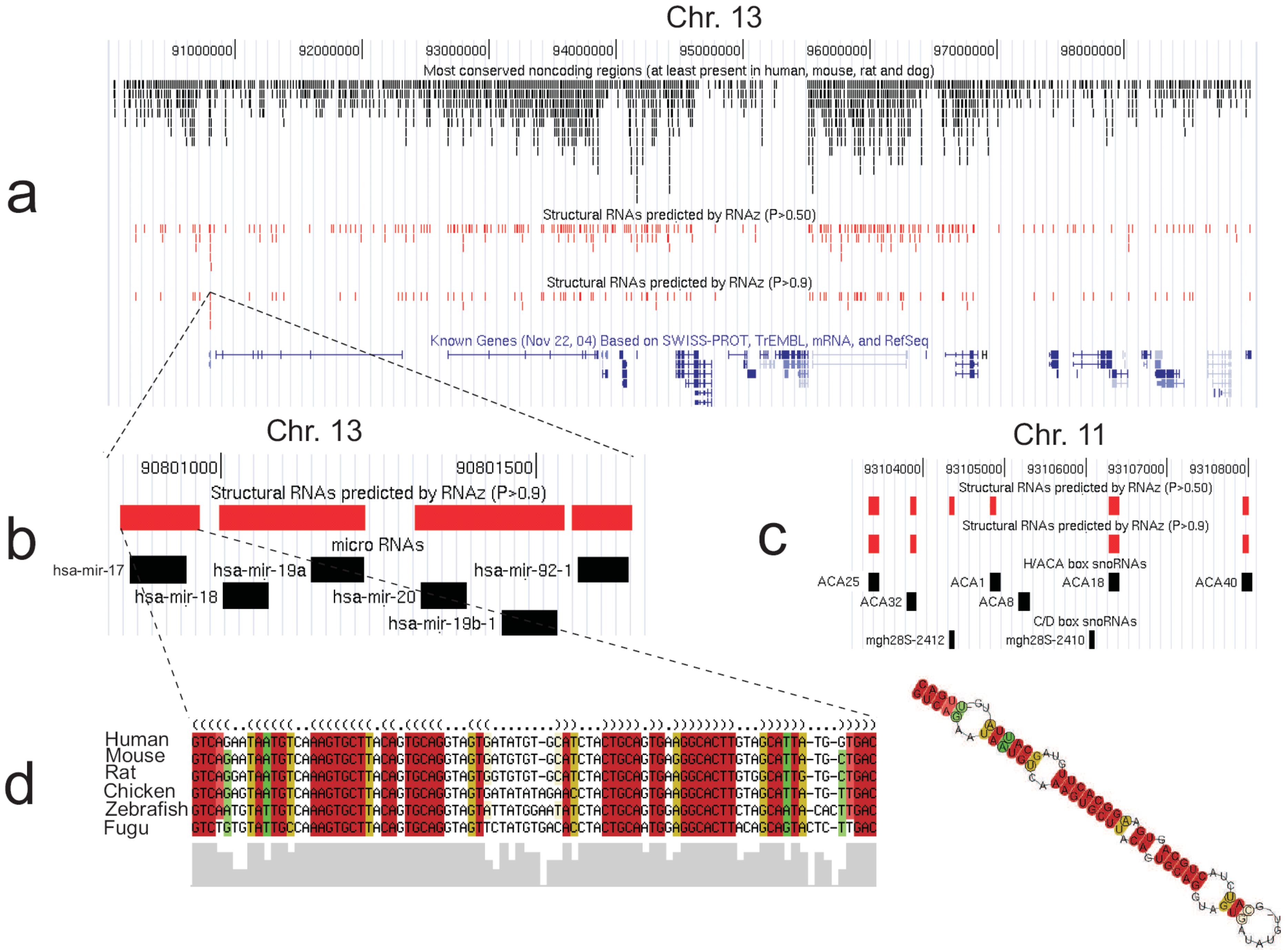
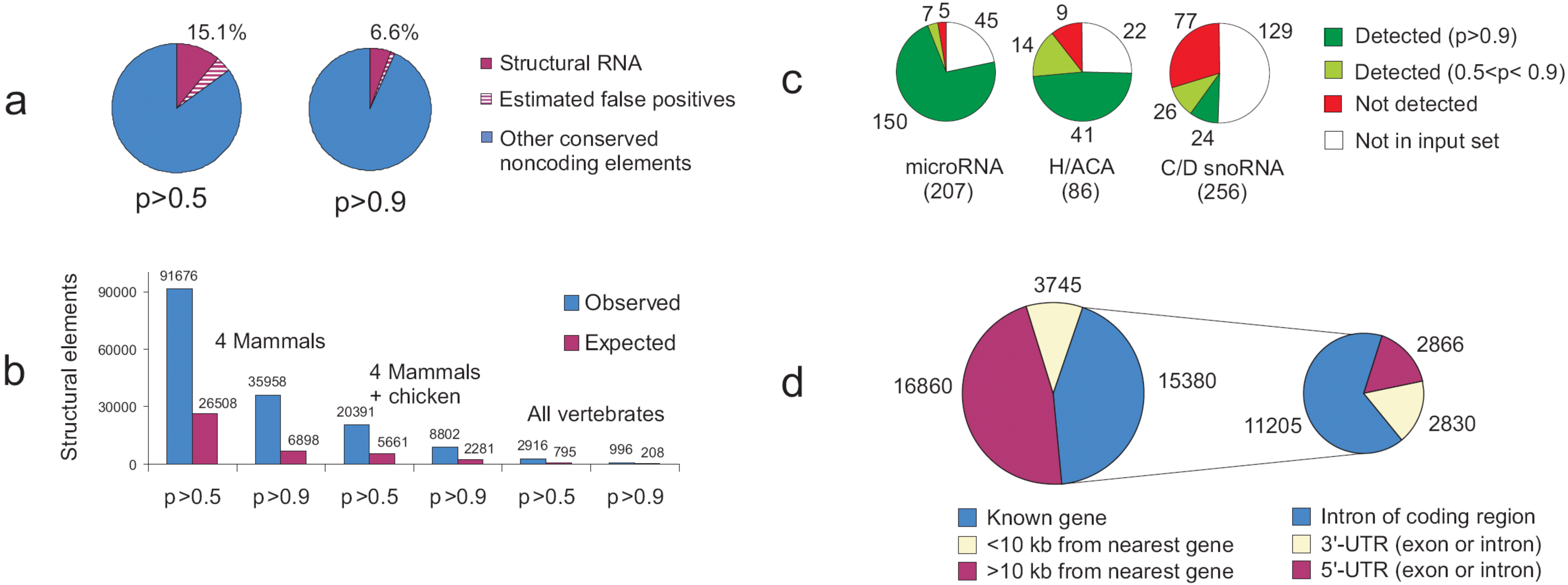


Table 1. Genomic coverage of filtering steps and phylogenetic conservation of ncRNA candidates.

	Genome Coverage		Alignments	RNAz hits $p > 0.9$		
	Size (MB)	Fraction (%)	Number	Size (MB)	Fraction of input (%)	Number
Human genome	3,095.02	100.00	–			
PhastCons most conserved	137.85	4.81	1,601,903			
without coding regions	110.04	3.84	1,291,385			
without alignments $< 50nt$	103.83	3.33	564,455			
Set 1: 4 Mammals	82.64	2.88	438,788	5.46	6.62	35,985
Set 2: + Chicken	24.00	0.85	104,266	1.34	5.50	8,802
Set 3: + Fugu or zebrafish	6.86	0.24	30,896	0.14	2.03	996


a
Chr. 13
91000000
92000000
93000000
94000000
95000000
96000000
97000000
98000000
Most conserved noncoding regions (at least present in human, mouse, rat and dog)
Structural RNAs predicted by RNAz (P>0.50)
Structural RNAs predicted by RNAz (P>0.9)
Known Genes (Nov 22, 04) Based on SWISS-PROT, TrEMBL, mRNA, and RefSeq
b
Chr. 13
90801000
90801500
Structural RNAs predicted by RNAz (P>0.9)
micro RNAs
hsa-mir-17
hsa-mir-18
hsa-mir-19a
hsa-mir-20
hsa-mir-19b-1
hsa-mir-92-1
c
Chr. 11
93104000
93105000
93106000
93107000
93108000
Structural RNAs predicted by RNAz (P>0.50)
Structural RNAs predicted by RNAz (P>0.9)
H/ACA box snoRNAs
ACA25
ACA32
ACA1
ACA8
ACA18
ACA40
C/D box snoRNAs
mgh28S-2412
mgh28S-2410
d
(((((..((((((..((((((((.((.(((((...(((.........)))...))))).)).))))))))...))))))....)))))
Human GTCAGAATAATGTCAAAGTGCTTACAGTGCAGGTAGTGATATGT-GCATCTACTGCAGTGAAGGCACTTGTAGCATTA-TG-GTGAC
Mouse GTCAGAATAATGTCAAAGTGCTTACAGTGCAGGTAGTGATGTGT-GCATCTACTGCAGTGAGGGCACTTGTAGCATTA-TG-CTGAC
Rat GTCAGGATAATGTCAAAGTGCTTACAGTGCAGGTAGTGGTGTGT-GCATCTACTGCAGTGAAGGCACTTGTGGCATTG-TG-CTGAC
Chicken GTCAGAGTAATGTCAAAGTGCTTACAGTGCAGGTAGTGATATATAGAACCTACTGCAGTGAAGGCACTTGTAGCATTA-TG-TTGAC
Zebrafish GTCAATGTATTGTCAAAGTGCTTACAGTGCAGGTAGTATTATGGAATATCTACTGCAGTGGAGGCACTTCTAGCAATA-CACTTGAC
Fugu GTCTGTGTATTGCCAAAGTGCTTACAGTGCAGGTAGTTCTATGTGACACCTACTGCAATGGAGGCACTTACAGCAGTACTC-TTGAC


a
15.1%
6.6%
p>0.5
p>0.9
Structural RNA
Estimated false positives
Other conserved noncoding elements
b
Structural elements
90000
60000
30000
0
91676
26508
35958
6898
20391
5661
8802
2281
2916
795
996
208
4 Mammals
4 Mammals + chicken
All vertebrates
Observed
Expected
p>0.5
p>0.9
p>0.5
p>0.9
p>0.5
p>0.9
c
7
5
45
150
microRNA (207)
9
22
14
41
H/ACA (86)
77
129
26
24
C/D snoRNA (256)
Detected (p>0.9)
Detected (0.5<p< 0.9)
Not detected
Not in input set
d
3745
16860
15380
2866
11205
2830
Known gene
<10 kb from nearest gene
>10 kb from nearest gene
Intron of coding region
3'-UTR (exon or intron)
5'-UTR (exon or intron)