Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA structure

Elfar Þórarinsson

February 2006

It’s interesting to note that:

- Approximately half of the ~3.000 million nucleotides in the human genome are masked by repeats.
- Roughly two thirds of the remaining nucleotides can be aligned with mouse (http://genome.ucsc.edu/).
- About one third of the whole, non-repeat, human genome is unalignable with the mouse.

Transcribed ncRNAs

- A very crude estimate, based on the transcriptional maps, for ten chromosomes, by Cheng et al. (2005), implies that roughly 32% of the human genome is transcribed.
- The majority of these transcripts (60-84%) don’t overlap with exons of known protein coding genes, indicating a considerable amount of ncRNAs to be found and annotated.
- It has also been implied that a large fraction of the mouse genome is non-coding (Suzuki and Hayashizaki, 2004, FANTOM consortium, 2005).

The question we seek to answer is:

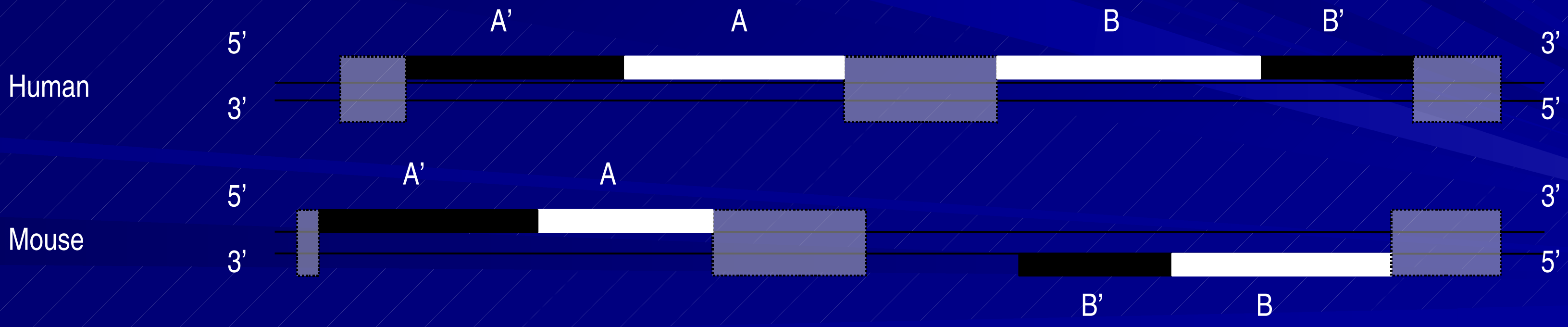
Are there places in the assumed non-conserved regions of the mammals that have evolutionary constraints on maintaining their structure?

FOLDALIGN

- An effective implementation of the Sankoff (1985) algorithm
- Aligns structure and sequence
- Locally aligns two sequences using dynamic programming to fill out a 4-dimensional matrix
- Scores using energy and sequence similarity parameters
- Calulates P-Score in an BLAST-like manner
- Described in Havgaard et al. 2005 and webserver accessible at http://foldalign.kvl.dk

Human Vs. Mouse

- How to limit the search space due to computational complexity?
- Scan unalignable sequence pairs that lie adjacent to a matching alignment



Processing Foldalign output

- Initial Filtering
  - Remove sequences with less than 40% of nucleotides involved in basepairing
  - Remove sequences shorter than 60 nt
- Secondary filtering
  - Randomize the the pairs and run FOLDALIGN on these
  - Compare the original pairs to the randomized pairs

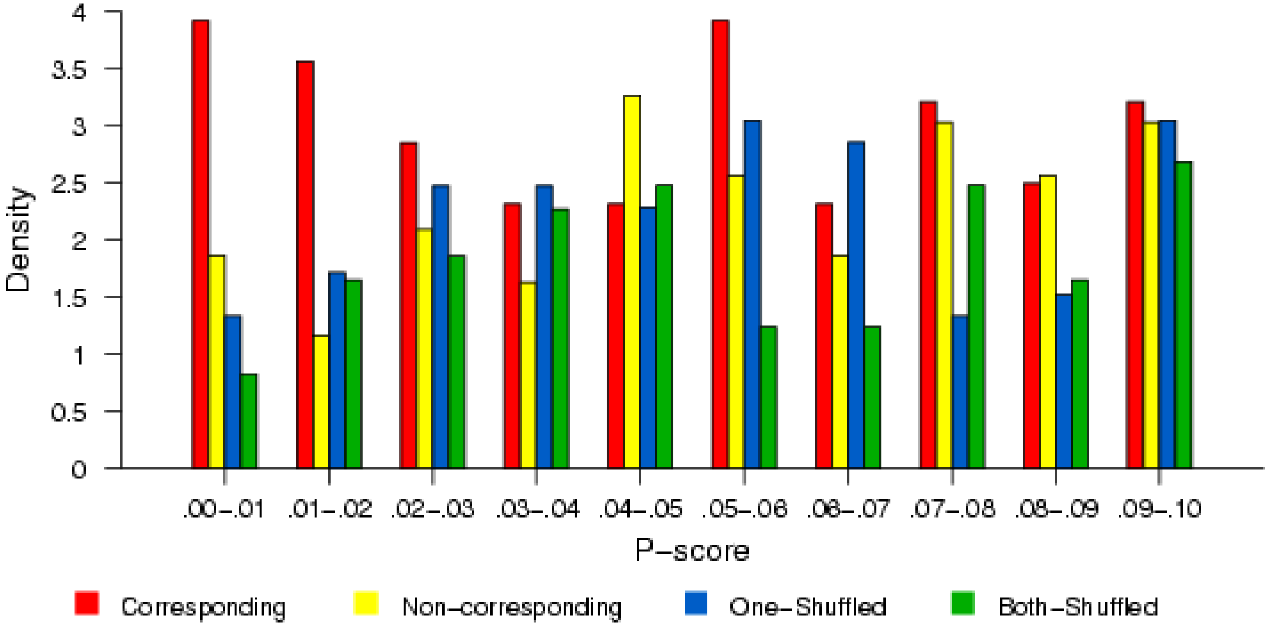
Human Vs. Mouse - Results

- Roughly 100.000 pairs, thereof 37.000 belong to the 10 chromosomes that have transcriptional maps
- Chromosome 20 was chosen as a model chromosome
- Chromosome 20 contains 2 x 3905 pairs
- 2260 alignments with length > 60 and basepairing > 40
- Half of these overlap transfrags

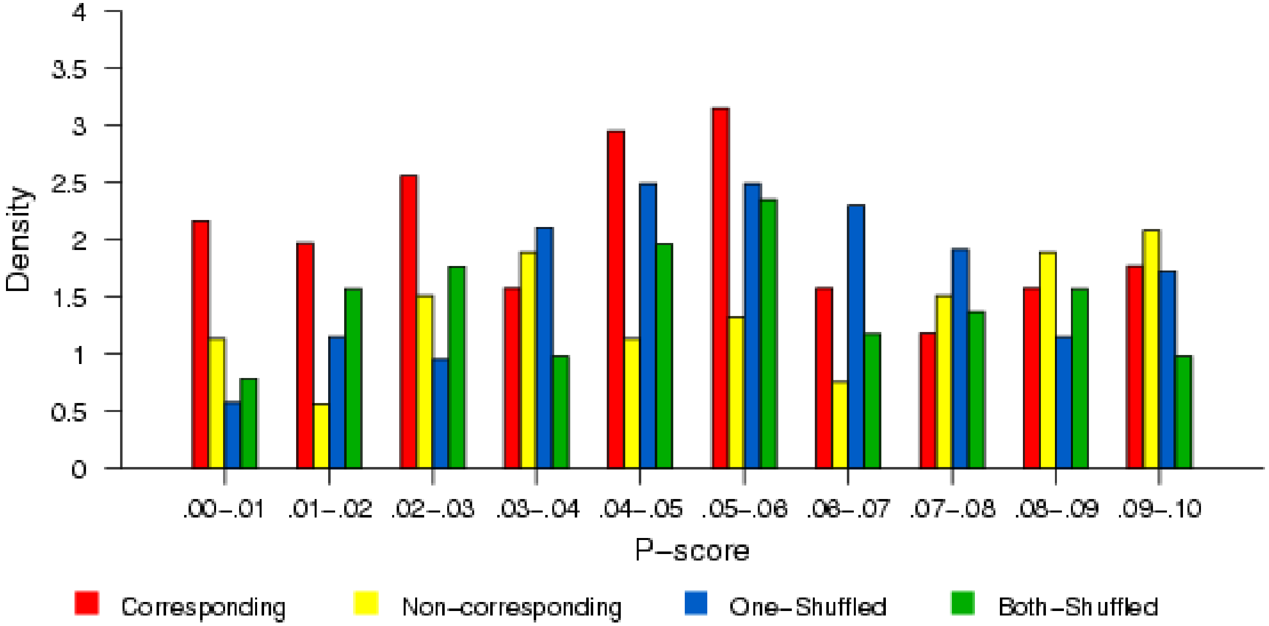
Randomizations

- We performed 3 different randomizations on chromosome 20
  - Shuffle both sequences in the pairs, maintaining dinucletide composition
  - Shuffle either only the human or only the mouse sequences, maintaining dinucletide composition
  - Randomize the pairs, i.e. scan probable non-corresponding pairs

Transfrag-Overlapping

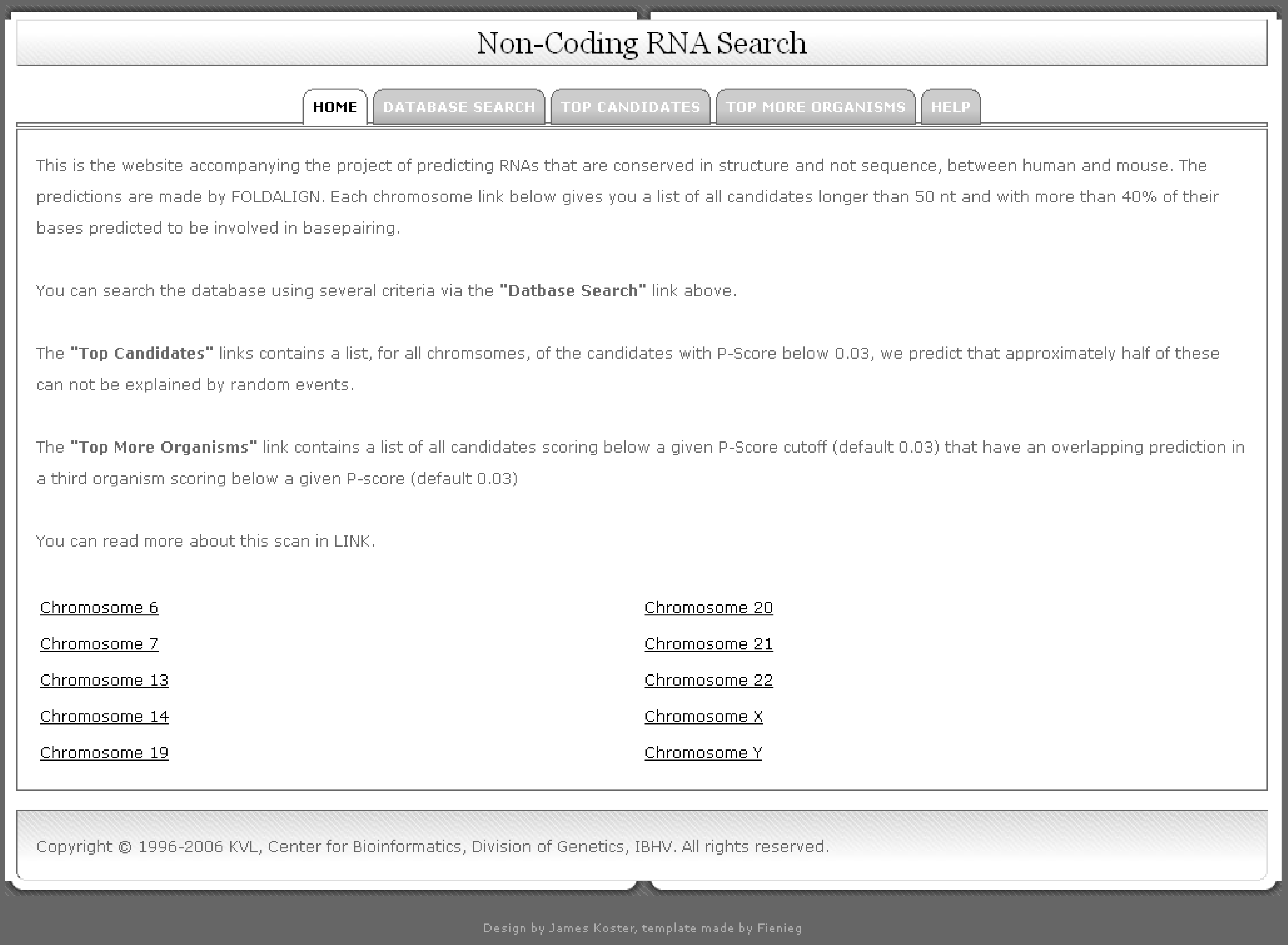


Non-Transfrag-Overlap



Candidate information

- Gather much information, i.e. transfrag info, known genes, predicted genes, structure, ESTs, FANTOM etc. etc.
- The information is kept in a MySQL database
- The database can be accessed via a PHP frontend
- The database also contains and generates “on the fly” .bed files which can be viewed in the UCSC genome browser



Non-Coding RNA Search

This is the website accompanying the project of predicting RNAs that are conserved in structure and not sequence, between human and mouse. The predictions are made by FOLDALIGN. Each chromosome link below gives you a list of all candidates longer than 50 nt and with more than 40% of their bases predicted to be involved in basepairing.

You can search the database using several criteria via the **"Datbase Search"** link above.

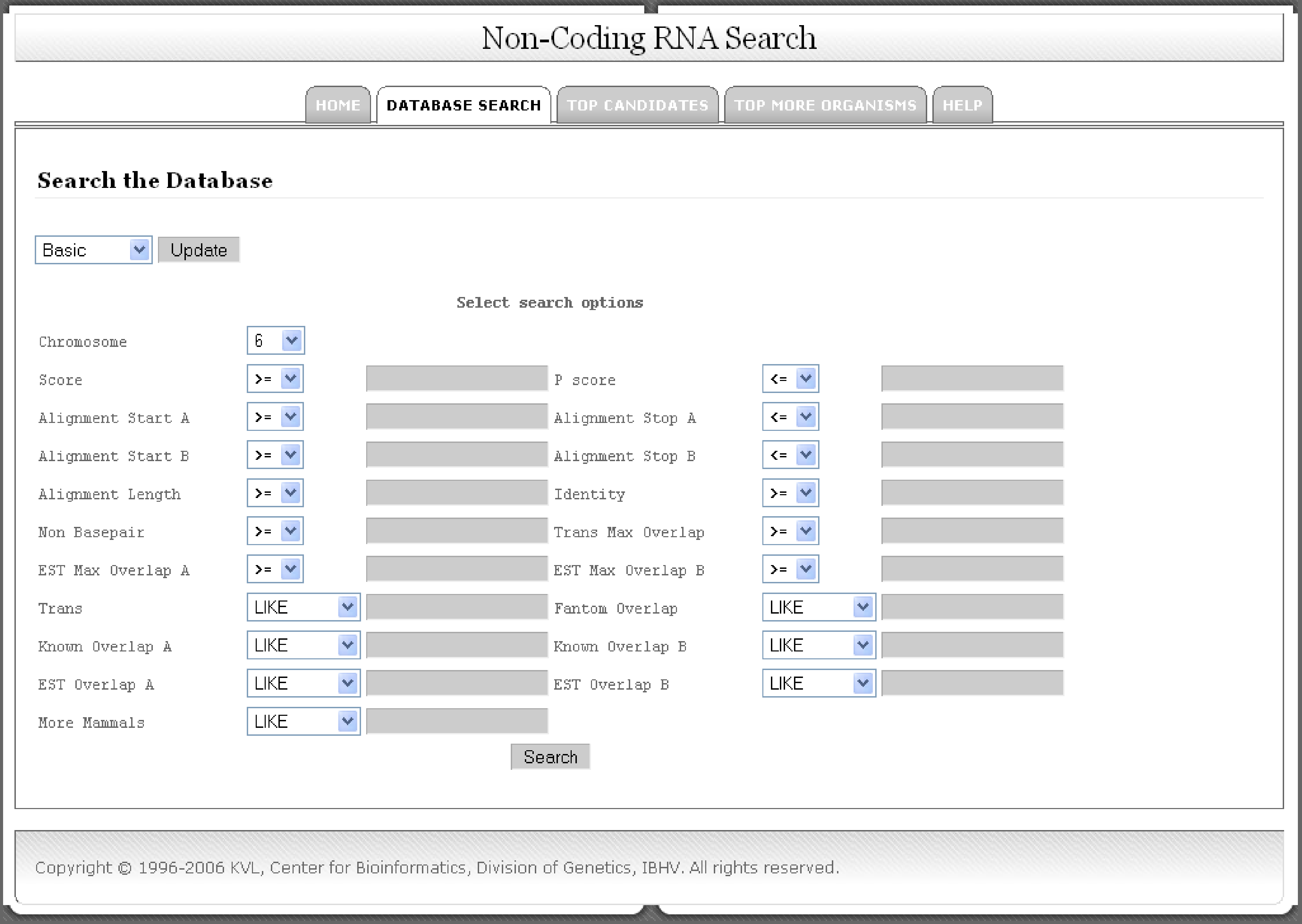
The **"Top Candidates"** links contains a list, for all chromsomes, of the candidates with P-Score below 0.03, we predict that approximately half of these can not be explained by random events.

The **"Top More Organisms"** link contains a list of all candidates scoring below a given P-Score cutoff (default 0.03) that have an overlapping prediction in a third organism scoring below a given P-score (default 0.03)

You can read more about this scan in LINK.

Chromosome 6
Chromosome 7
Chromosome 13
Chromosome 14
Chromosome 19
Chromosome 20
Chromosome 21
Chromosome 22
Chromosome X
Chromosome Y

Design by James Koster, template made by Fienieg



Non-Coding RNA Search

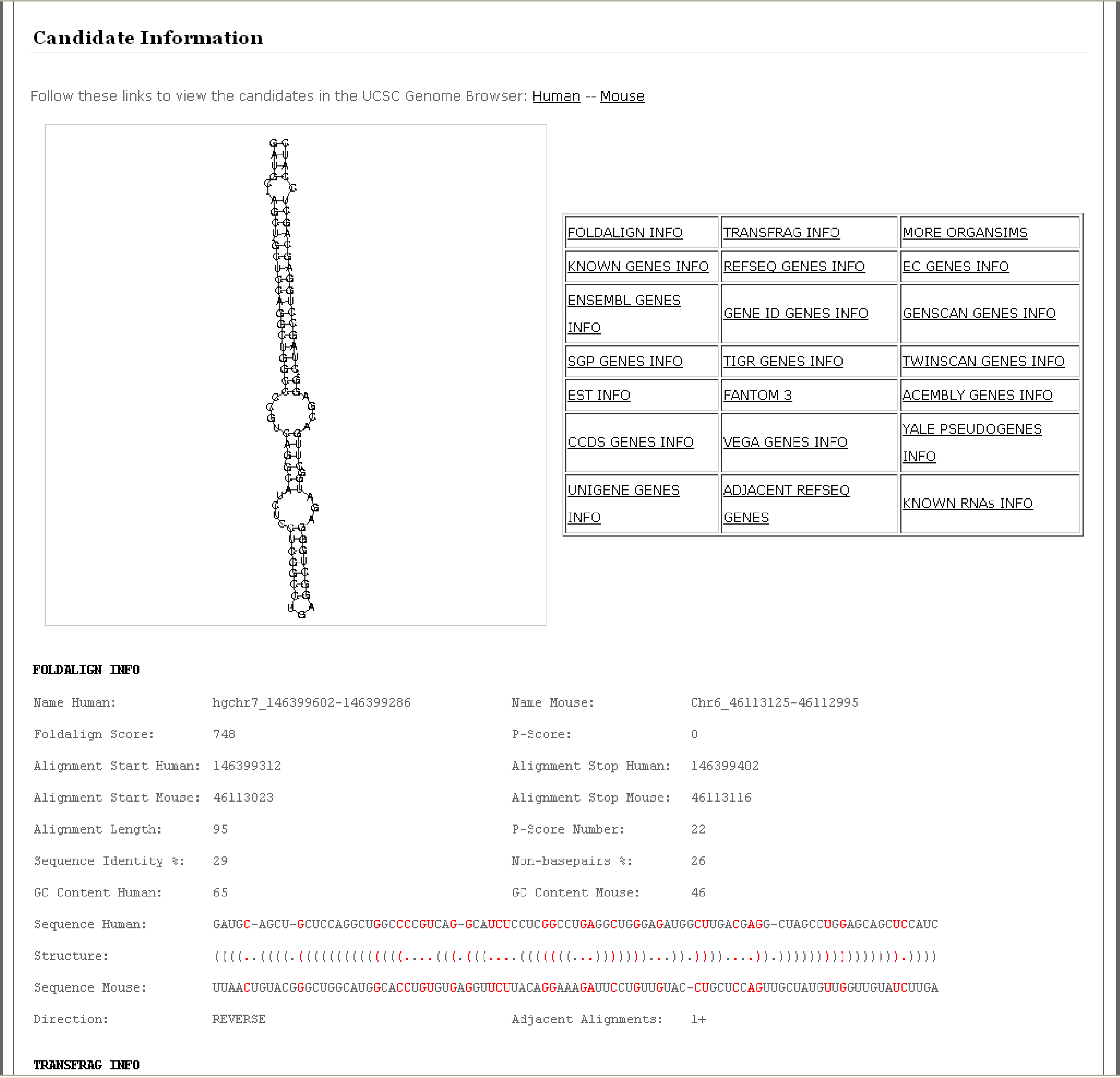
Search the Database

Basic

Select search options

Chromosome	6				
Score	>=		P score	<=	
Alignment Start A	>=		Alignment Stop A	<=	
Alignment Start B	>=		Alignment Stop B	<=	
Alignment Length	>=		Identity	>=	
Non Basepair	>=		Trans Max Overlap	>=	
EST Max Overlap A	>=		EST Max Overlap B	>=	
Trans	LIKE		Fantom Overlap	LIKE	
Known Overlap A	LIKE		Known Overlap B	LIKE	
EST Overlap A	LIKE		EST Overlap B	LIKE	
More Mammals	LIKE				

Design by James Koster, template made by Fienieg



Candidate Information

Follow these links to view the candidates in the UCSC Genome Browser: Human -- Mouse

FOLDALIGN INFO	TRANSFRAG INFO	MORE ORGANSIMS
KNOWN GENES INFO	REFSEQ GENES INFO	EC GENES INFO
ENSEMBL GENES INFO	GENE ID GENES INFO	GENSCAN GENES INFO
SGP GENES INFO	TIGR GENES INFO	TWINSCAN GENES INFO
EST INFO	FANTOM 3	ACEMBLY GENES INFO
CCDS GENES INFO	VEGA GENES INFO	YALE PSEUDOGENES INFO
UNIGENE GENES INFO	ADJACENT REFSEQ GENES	KNOWN RNAs INFO

FOLDALIGN INFO

Name Human:	hgchr7_146399602-146399286	Name Mouse:	Chr6_46113125-46112995
Foldalign Score:	748	P-Score:	0
Alignment Start Human:	146399312	Alignment Stop Human:	146399402
Alignment Start Mouse:	46113023	Alignment Stop Mouse:	46113116
Alignment Length:	95	P-Score Number:	22
Sequence Identity %:	29	Non-basepairs %:	26
GC Content Human:	65	GC Content Mouse:	46

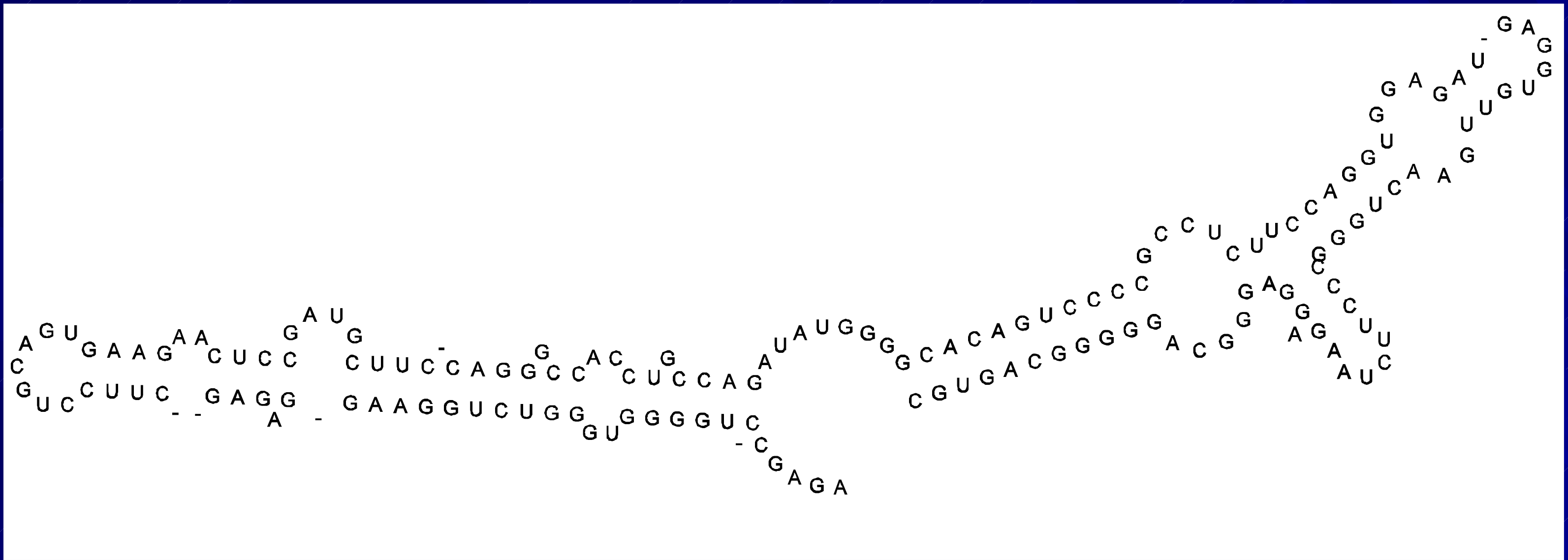
```
Sequence Human:  GAUGC-AGCU-GCUCCAGGCUGGCCCCGUCAG-GCAUCUCCUCGGCCUGAGGCUGGGAGAUGGCUUGACGAGG-CUAGCCUGGAGCAGCUCCAUC

Structure:       ((((..((((.((((((((((((((....(((.(((....(((((((...)))))))...)).))))....)).)))))))))))))))).))))

Sequence Mouse:  UUAACUGUACGGGCUGGCAUGGCACCUGUGUGAGGUUCUUACAGGAAAGAUUCCUGUUGUAC-CUGCUCCAGUUGCUAUGUUGGUUGUAUCUUGA
```

Direction: REVERSE

Adjacent Alignments: 1+

TRANSFRAG INFO

Example Candidate



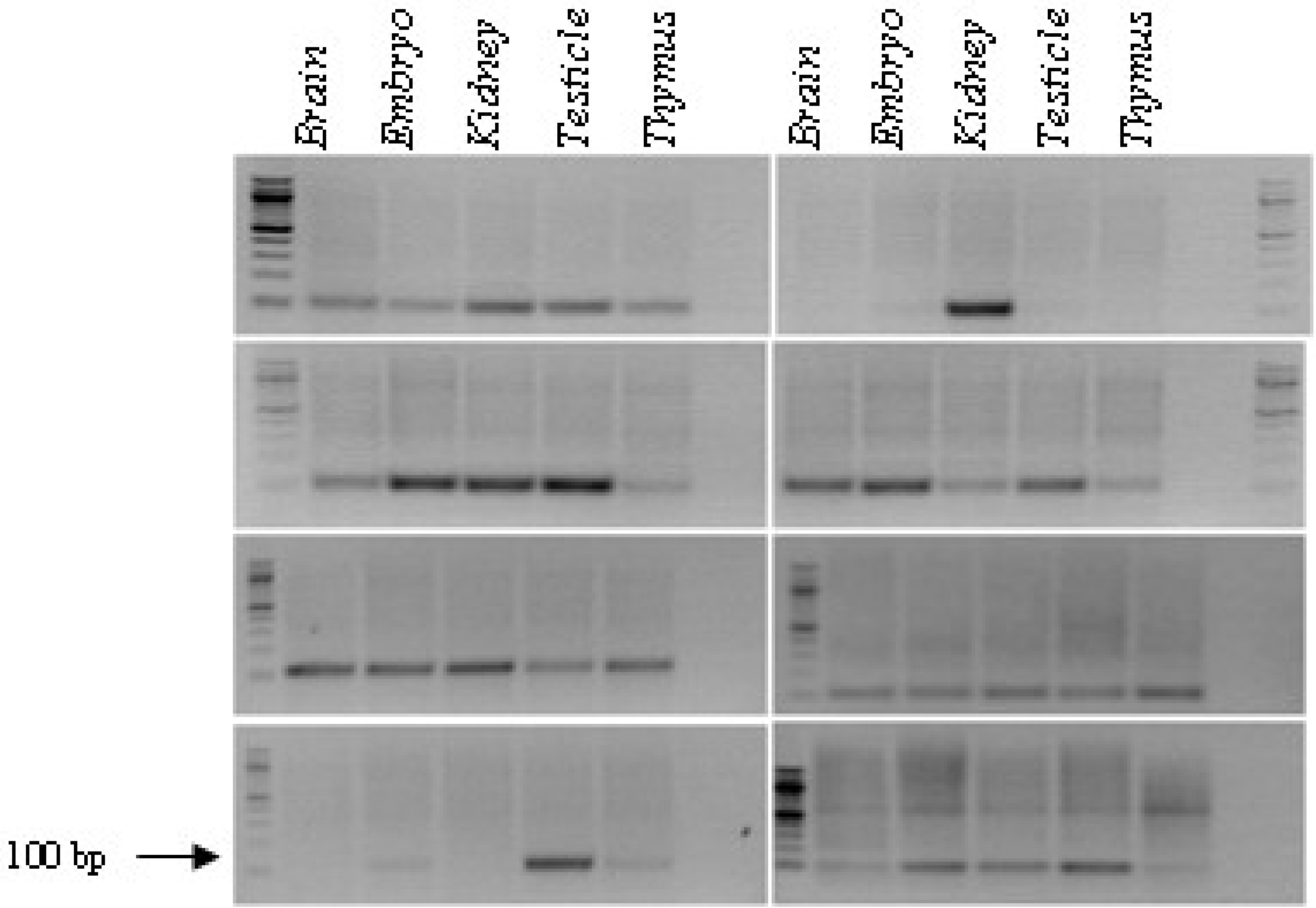
```
AGAGCC-UGGGGUGGGUCUGGAAG-GAGAG--CUUCCUGCAGUGAAGAACUCCGAUGCUUC-CAGGGCCACCUGCCAGAUAUGGGCACAGU
.....(.(((((..((((((((((.(.(((..((((.......))))..))))....)))).)))).))..)).)))).......((((.((
AGAGCUACAGGGAAGCGGGGCUGGCUAGUCCCUUGGCUG--GCCUGACUGACAUACAUCAGACCCUGGCUGCC-CUGGGUAUAG-AAGCAAC
```

```
CCCCGCCUCUUCCAGGUGGAGAU-GAGGUGUUGAACUGGGCCCUUCUAAGAGGAGGGCAGGGGGCAGUGC
((((....((.((.(((...(((......)))..))).))((((....)).))))....)))))).))))
UCUGAACUUGUCCAUGCU-AGGCAGAUGUGCCUAGCAGGGUUCUAACAGGAAGCAGCUGCAGGGUGGCUU
```

Experiments

- We have performed RT-PCR on total mouse RNA using oligo dT and random hexamers, and then gene-specific primers
- 32/36 top-scoring transfrag-overlapping candidates were verified
- 7/9 top-scoring non-transfrag-overlapping candidates were verified
- We performed Northern blotting on 12 of the 36 candidates, 4 gave positive results

Some transfrag candidates



Conclusion

- Our findings suggest that there are corresponding regions between human and mouse which contain orthologous expressed non-coding RNA sequences not alignable in primary sequence
- In human we estimate 4000 < 0.03 candidates, half of these 4000 we cannot explain by random events

Acknowledgements

Milena Sawera

Jakob Havgaard

Merete Fredholm

Jan Gorodkin

Danish research councils SJVF and STF

Danish Center for Scientific Computing