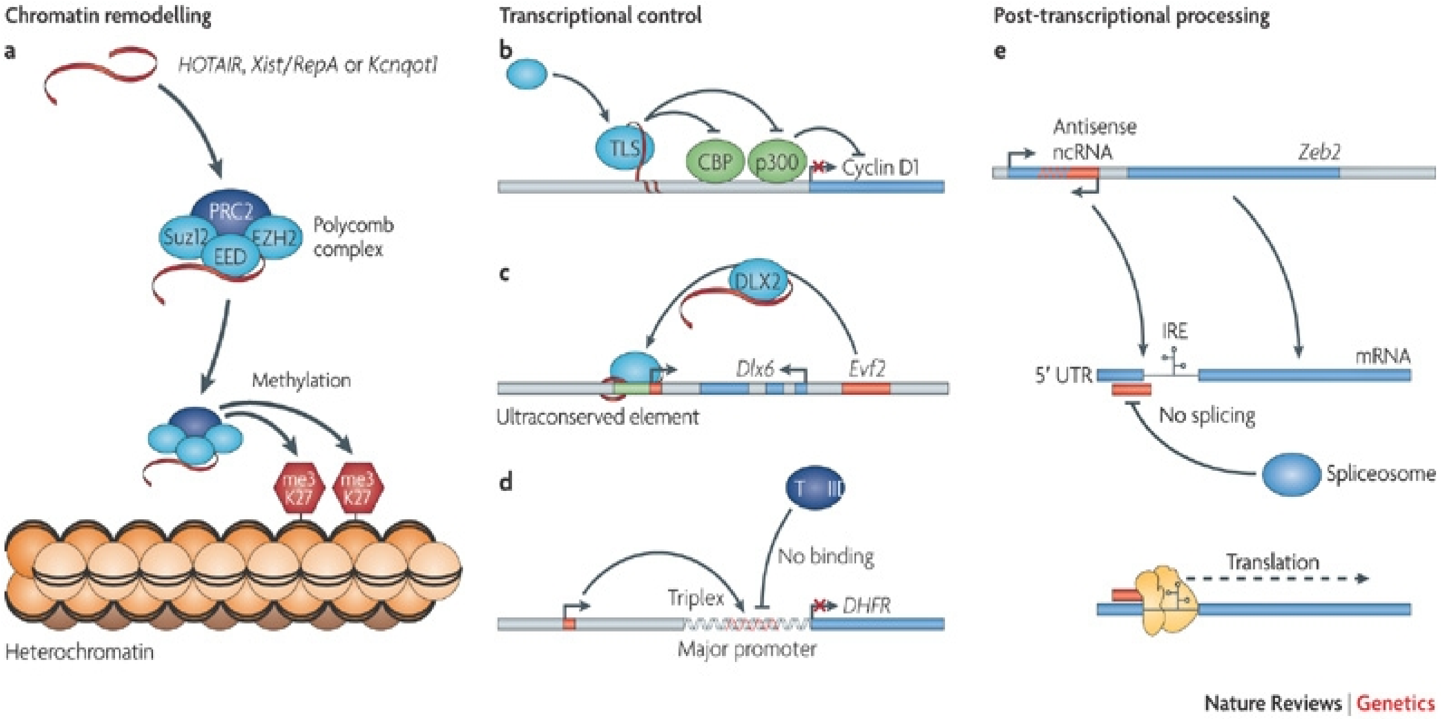
Multiple sequence alignments of long non-coding RNAs

Ilinca Tudose,
under the coordination of Dr. Dominic Rose

29th July 2011

Long non-coding RNAs



[Mercer et al., Nature Genetics, 2008]

There is only scarce knowledge about lncRNAs.

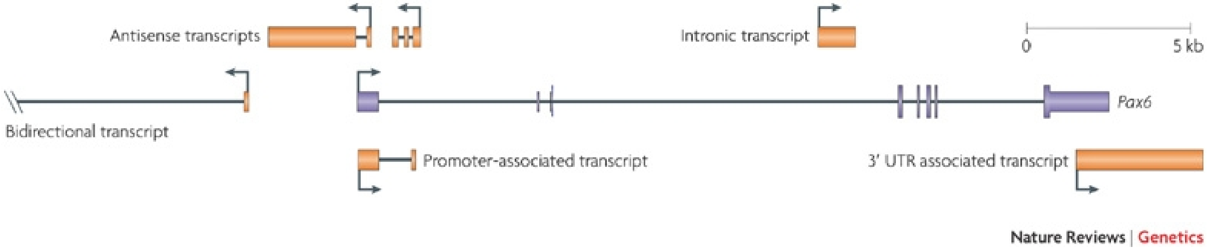
Medical significance : due to their ability to regulate associated protein coding genes.

LncRNA – specificities

Long. Generally any lncRNAs which is over 200 bases long , but often a lot longer.

General low sequence conservation.

Very complex - they often overlap or occur between multiple coding and non-coding sequences


Antisense transcripts
Intronic transcript
0
5 kb
Pax6
Bidirectional transcript
Promoter-associated transcript
3' UTR associated transcript
Nature Reviews | Genetics

[Mercer et al., Nature Genetics, 2008]

Aim Of The Project

Study the performance of leading existing alignment tools and a new approach in the context of lncRNAs.

Motivation: many subsequent analyses regarding lncRNAs require valid alignment as input

Input: set of 93 lncRNAs (www.lncRNAdb.org)

Steps:

- Develop our own “alignment tool”
- Try existing tools on the same set
- Score and compare results

Alignment approach

Input: one human lncRNA sequence

Output: multiple alignment with homologous RNAs in as many species as possible

Approach:

(1) Search for homologous sequences in other species

(2) Build a set of the best finds

(3) Align the set of sequences

Alignment approach - details

- BLAT the human lncRNA against different species

  (out of a set of 19 species) and parse result page.
- Add new sequence to the set and align (using MUSCLE)
- Decide whether to keep the new sequence or not

  (given the score)

Alignment approach – details (cont)

Parameters (constraints):

- Intron length
- Score
- Number of gaps

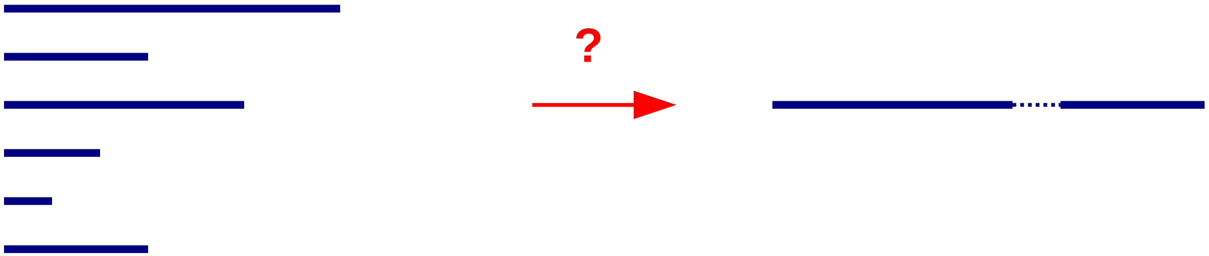
Sounds easy, right?
Well, it always does but ...

Try & fail

Using the stand alone BLAT version:

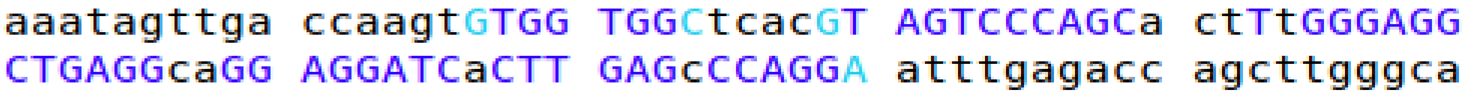
difficult to recover the sequence from the reported blocks

(requires exhaustive post-processing)



Using the online BLAT tool:

Abundant creativity of the UCSC Genome web developers



```
aaatagttga ccaagtGTGG TGGCtcacGT AGTCCCAGCa ctTtGGGAGG
CTGAGGcaGG AGGATCaCTT GAGcCCAGGA atttgagacc agcttgggca
```

Comparison

Our approach

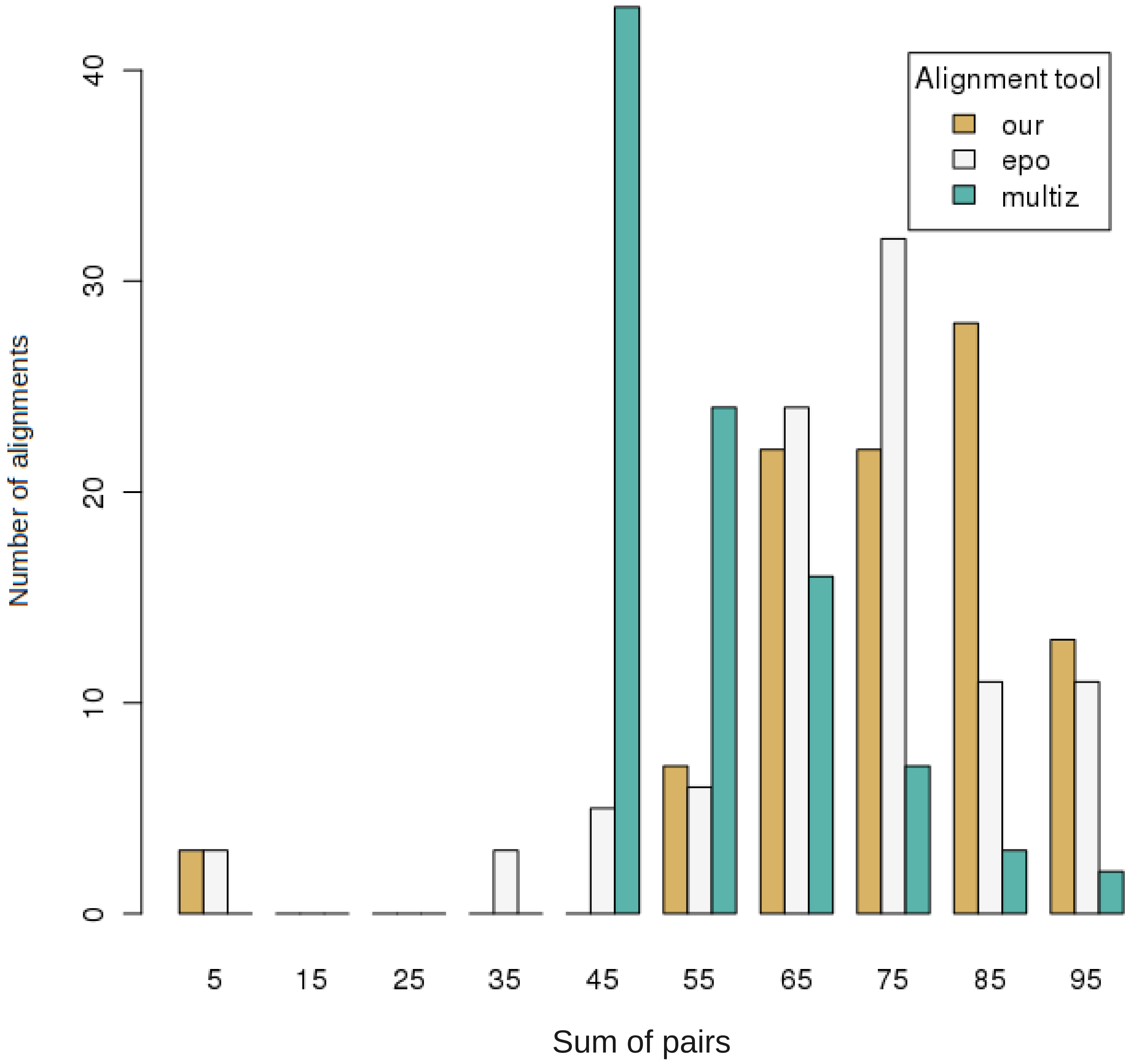
vs.

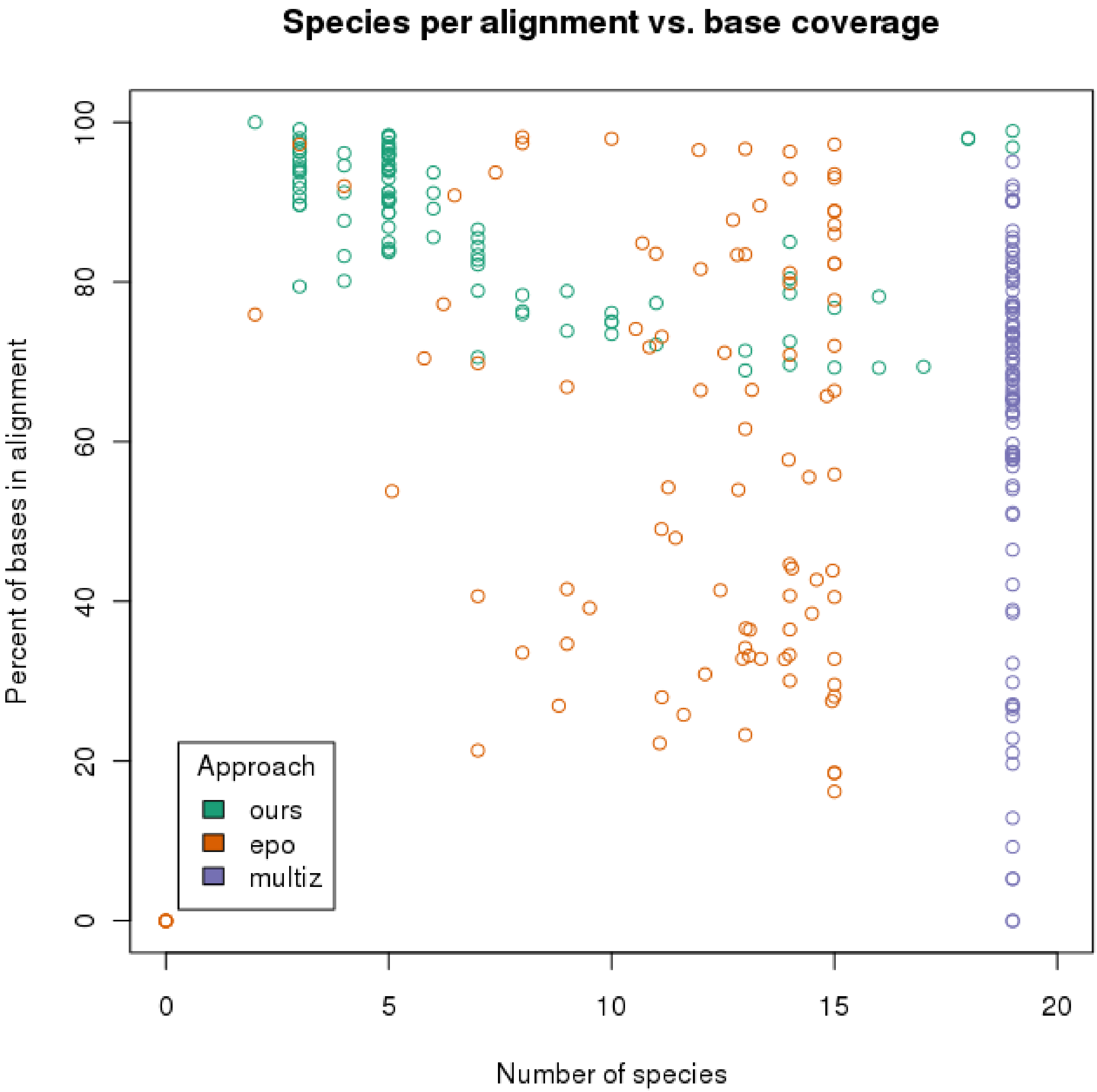
UCSC/Galaxy multiz alignments

vs.

Ensembl epo alignments

Sum-of-pairs scores (normalized) for the 3 alignment approaches




Species per alignment vs. base coverage
Percent of bases in alignment
0
20
40
60
80
100
0
5
10
15
20
Number of species
Approach
🟩 ours
🟧 epo
🟪 multiz

Comparison table

	our	multiz	epo
Loss of synteny	no	?	yes
Avg. # of sequences per alignment	7.1	19	11.6
Avg. sum-of-scores per alignment	77.8	56.1	68.5
Avg. % of bases in an alignment	87.5	62.1	57.3
mature mRNA alignments	no	yes	yes

Conclusions

Our approach:

Good scores but misses a lot of good matches
Currently limited by BLAT
Not very fragmented, no loss of synteny
Alignment of whole RNAs

Multiz:

Good balance score – number of sequences
Difficult to recover the alignment (from several exons)
Not enough information about the sequences in the alignment (i.e. check synteny lost?)

Ensemble:

Good balance score – number of sequences
Difficult to recover the alignment (from several exons)
Synteny lost

Further work

Play around with the treshhold parameters

Use the local BLAT, adjust the parameters
(especially the windowsize)

First build the whole set of “homologous” sequences,
then decide which to keep

Thank you!

References

Long non-coding RNAs: insights into functions, Nature Reviews Genetics 10, 155-159 (March 2009), Tim R. Mercer, Marcel E. Dinger & John S. Mattick

http://genome.ucsc.edu/

http://main.g2.bx.psu.edu/

http://www.ensembl.org/index.html

http://www.ebi.ac.uk/Tools/msa/muscle/

www.lncRNAdb.org