Detecting Structural Elements of lincRNAs using RNAz

Kaswara Kraibooj, Dominic Rose

Outline

1. Motivation: what is the project about?
2. Workflow, tools and methods
3. Results
4. Discussion

Outline

1. Motivation: what is the project about?

2. Work flow and tools and methods

3. Results

4. Discussion

Motivation: Input

- Dataset: 8195 transcripts of long intergenic non-coding RNAs of hg19 as a BED file

- Long intergenic non-coding (lincRNAs):
  - Long: length > 200 bp
  - Intergenic: stretches between the genes
  - Non-coding: do not code proteins

Motivation: Intended output

- Prediction of conserved (secondary) structural elements of lincRNAs using RNAz (classified as functional)

- Detection of common secondary structure motifs of the predicted elements using RNAclust

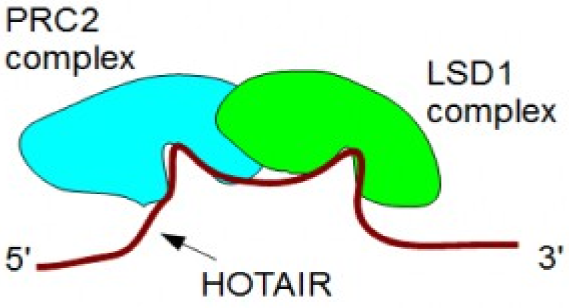
Motivation: LincRNAs (important?)

- Form the vast majority of RNA transcripts

- Regulate important biological processes in the cell
  - Example is HOTAIR

Motivation: LincRNAs example (HOTAIR)

- Cancer lincRNA (for HOX antisense intergenic RNA)
- Belongs to chromosome 12 (human genome)
- Interacts with two protein complexes together (LSD1, PRC2) to target genomic regions or genes (chromosome 2)
- Helps regulate immune response, cancer growth and production of cells.



Motivation: How (principle) ?

- Problem: the primary sequence of non coding RNAs does not have the same features as protein coding RNAs such as start/stop codons.

- Solution: exploit secondary structure (the function of ncRNAs are deeply related with their secondary structures)

- Detect:
  - Stable secondary structure
  - Conserved secondary structure

Outline

1. Motivation: what is the project about?
2. **Workflow, tools and methods**
3. Results
4. Discussion

Analysis Workflow

1. Fetch alignments
2. Pre-process the alignments (windows)
3. Predict the hits (windows)
4. Cluster the hits (windows) to loci
5. Estimate false discovery rate (FDR)
6. Find clusters of loci which share secondary structure motifs (tree)
7. Visualize the tree

1. Fetch alignments

- Using the free public server 'Galaxy'

- Two screens:
    1. Stitch gene blocks given a set of coding exon intervals (screen1)
    2. Extract MAF blocks given a set of genomic intervals (screen2)

2. Pre-process the alignments

- Using rnazWindow.pl:
    1. Get rid of gaps, repeats
    2. Split large alignments into smaller windows such as:
        - Length of one window 120 nt
        - Shift between the beginning of two successive windows is 40 nt
        - 120 and 40 result in optimal behavior of RNAz

3. Predict the hits (RNAz)

- Two independent measurements:
  1. Thermodynamical stability (z-score)
  2. Structural conversation Index (SCI)
- Classification: support vector machine learning (SVM) algorithm trained on a large number of well known ncRNA.
- Predicted sequences of probability value bigger than 0.5 are classified as functional

4. Cluster the hits

- It clusters the overlapping windows in one hit to one locus.

- Locus: the stretch on the overlapping windows which have one hit, from the beginning of the first window to the end of the last window.

5. Estimate false discovery rate (FDR)

- Statistical measurement of the error percentage of the predicted hits number

- FDR_1= $\frac{number\ of\ windows(shuffled)}{number\ of\ windows(original)}$

- FDR_2= $\frac{number\ of\ loci(shuffled)}{number\ of\ loci(original)}$

- FDR_3= $\frac{length\ of\ loci(shuffled)}{length\ of\ loci(original)}$

5. Estimate false discovery rate (FDR)

- How:
  1. Shuffle the windows (SISSIz)
  2. Run RNAz again but on the shuffled windows
  3. Calculate FDR

- Note: SISSIz tools do not change the alignment characteristics.

5. Detect common secondary structure motifs

- RNAclust.pl clusters the loci in order to discover shared secondary structure motifs.

- Its output is a tree:
    - its internal nodes are the clusters
    - its leaves are the loci sequences

6. Visualize the tree

- The tree resulting from RNAclust.pl can be visualized by iTOl and Soupviewer.

- Purpose: facilitate showing and studying the tree.

- iTOL has more visualizing abilities than Soupviewer does such as coloring according to colors defined by the user.

Outline

1. Motivation: what is the project about?
2. Work flow and tools and methods
3. **Results**
4. Discussion

Results: Comparison

Number of loci	screen1	screen2
(0.5)	2279	8158
(0.9)	877	2700

Loci of high reliability (0.9):

- 2700 loci are located in introns and exons, or span splice sites.
- 1823 (67.51%) are located in introns.
- 197 (7.29%) span splice sites.
- 680 (25.18%) are located in exons.

Results: Comparison

- Loci of high reliability (0.9):
  - 2700 loci are located in introns and exons, or span splice sites.
  - 1823 (67.51%) are located in introns.
  - 197 (7.29%) span splice sites.
  - 680 (25.18%) are located in exons.

- Observation: there are more hits in introns than hits in exons or/and splice sites.
- Possible interpretation: small ncRNAs are hosted by lncRNAs in their introns.

Results: FDR

- The average value of the FDRs resulting from the three used equations are:

	screen1	screen2
FDR (0.5)	0.30	0.60
FDR (0.9)	0.28	0.48

- FDR of screen1 is optimistic, whereas FDR of screen2 is pessimistic.

Results: FDR

- In screen1 (0.9): approximately 630 loci (72 %) show signals of stability and conservativity, and therefore are most likely functional of high reliability.

- In screen2 (0.9): more than 1400 loci (52%) show signals of stability and conservativity, and therefore are most likely functional of high reliability.

Results: FDR

- In screen1 (0.5): approximately 1595 loci (70 %) show signals of stability and conservativity, and therefore are most likely functional (optimistic).

- In screen2 (0.5): about 3263 loci (40 %) show signals of stability and conservativity, and therefore are most likely functional (pessimistic).

- Observation: FDR of screen2 is remarkably bigger than FDR in screen1
- possible interpretation: future work!

Results: Transcripts/screen1

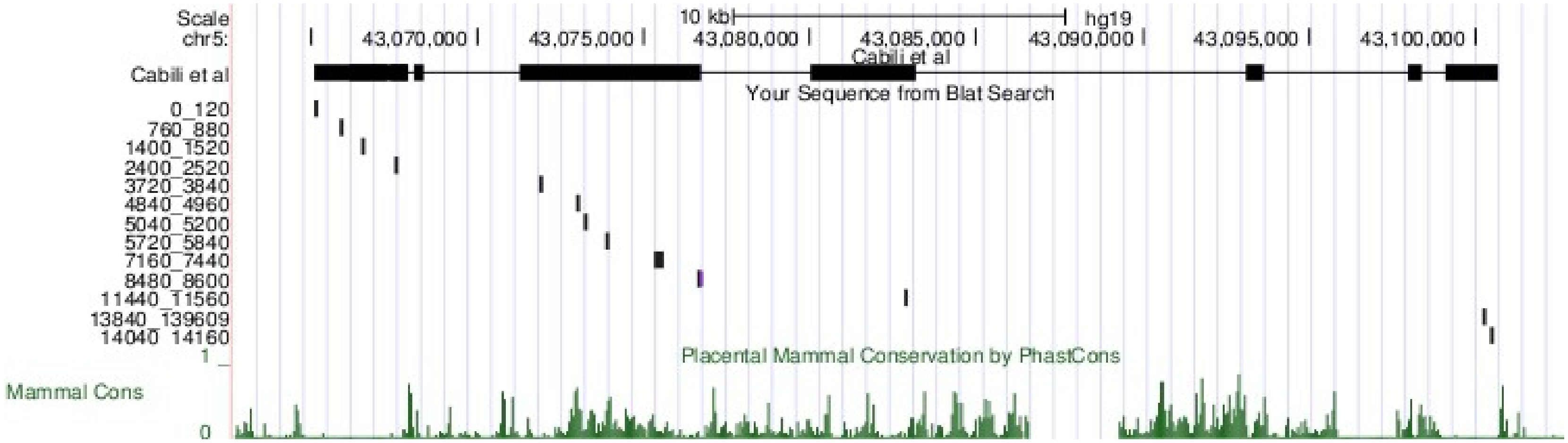
	Number of transcripts	Length of transcripts
Original	8195 (100%)	426642903 (100 %)
Alignments	99.75%	31%
Windows	66.21%	0.94%
Loci (0.5)	15.57%	0.076%
Loci (0.9)	7.78%	(0.027%)

- Observation: the total length of loci is too small
- Interpretation: the big loss of signals by aligning and windowing
- Conclusion: not sufficient

Results: 13 hits in one transcript

Number of transcripts	Number of hits
476	1
119	2
25	3
13	4
4	5
1	6
1	**13**

Results: 13 hits in one transcript (UCSC)



- Observation: loci are grouped in clusters.
- Possible interpretation: they might have a common function.

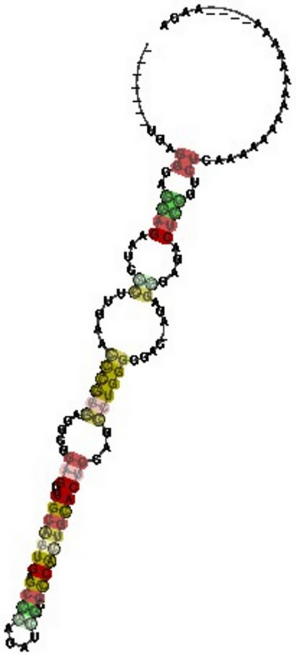
Results: Common secondary structure

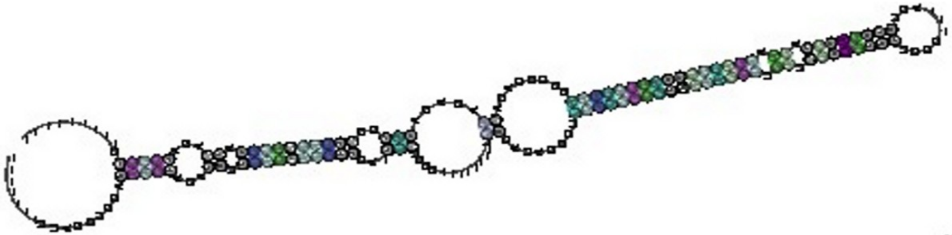
- Input: loci
- Output: clusters share common secondary structure motifs, represented by a hierarchical tree.

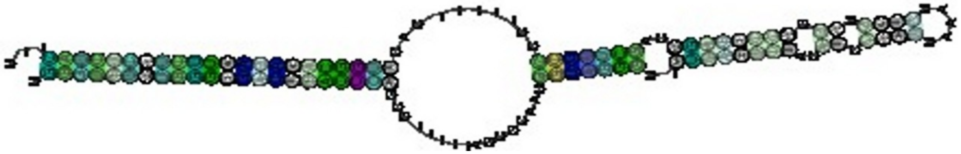
Node_ID	Number of loci	MFE	SCI
1284	14	-24.54	0.56
1568	10	-42.29	0.57
138	7	-39.15	0.77

- Observation: few loci clusters have shared secondary motifs.
- Possible biological meaning: the loci have few common functions.

Results: Common secondary structure

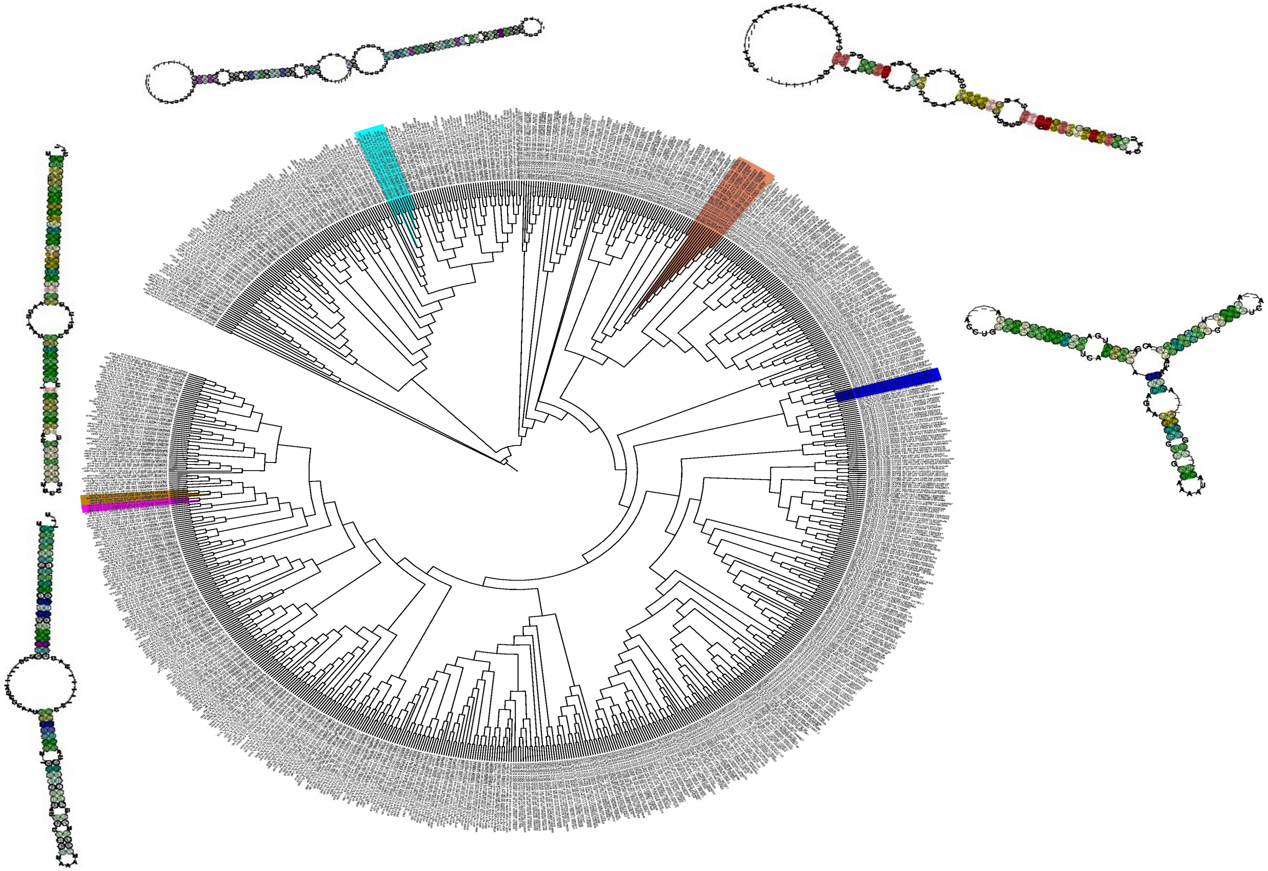






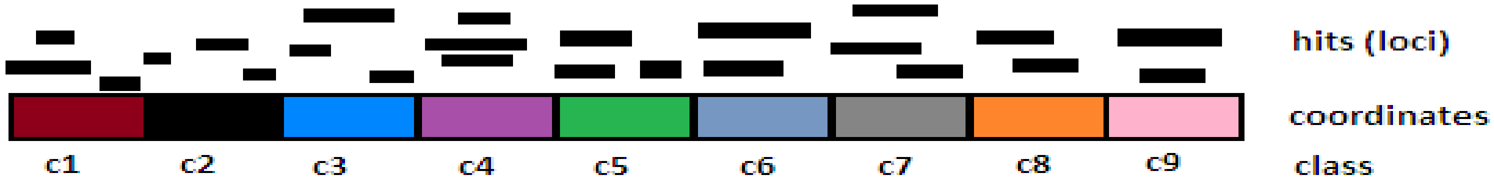
Node_ID	Number of loci	MFE	SCI
1284	14	-24.54	0.56
1568	10	-42.29	0.57
138	7	-39.15	0.77

Results: Common secondary structure (iTOL)

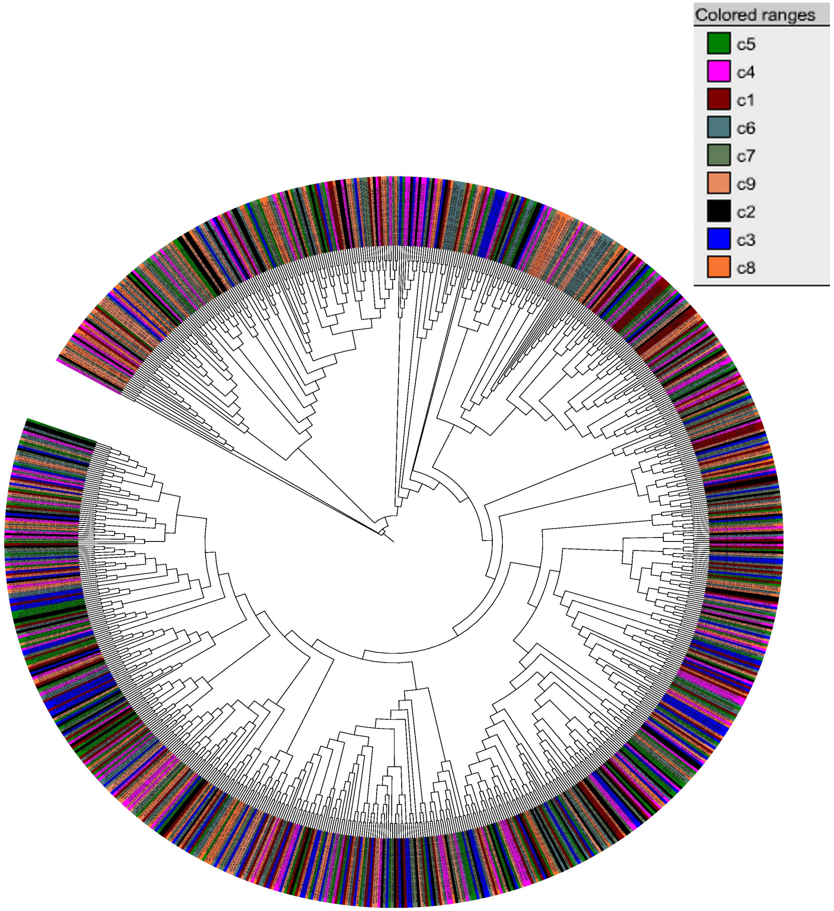


Results: Similarity of loci located in the same relative locations

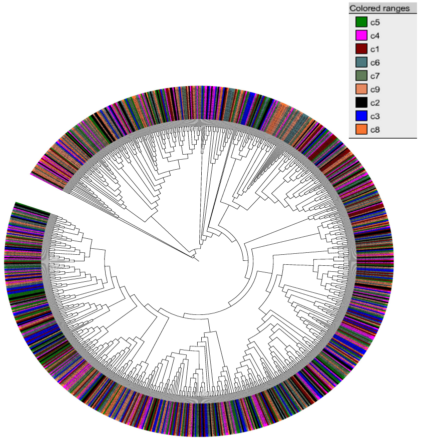
- The relative locations of the hits on its transcripts were segmented into nine locations (classes: c1, c2, c3, ..., c9).
- The hits that belong to one class have the same color.



Results: Similarity of loci located in the same relative locations (visualized by iTOL)



Results: Similarity of loci located in the same relative locations (visualized by iTOL)



- Observation: the colors are randomly distributed
- Conclusion: the hits that are located in the same segments of the same or different transcripts do not show common secondary structures.

Outline

1. Motivation: what is the project about?
2. Work flow and tools and methods
3. Results
4. **Discussion**

Discussion

- Interesting:
  - Most of the hits are located in introns. Why?
  - FDR of screen2 is bigger than FDR of screen1. Why?

- Negative:
  - The total length of the hits (loci) is too small for considering the analysis as comprehensive and sufficient one.
  - The number of clusters of the loci which have common secondary structure are few.
  - Similarity in secondary structure according to the relative locations are too sparse.

Discussion

- Positive:
    - Nevertheless, the 13 loci of chromosome 5 are very interesting and might together have biological functionality.

    - Additionally, the loci of one cluster, which share secondary structure motif, likely have the same function.

- Drawbacks:
    - The length of the aligned sequences are too short
    - Suggestion: new methods do not depend on alignments

Vielen Dank!