Annotating long non-coding RNAs in model and non-model organisms using a Random Forest strategy

FEELnc: FlExible Extraction of LncRNAs

Valentin Wucher
Roderic Guigó's Lab
Centre de Regulació Genòmica (CRG) - Barcelona - Spain

Séminaire MIAT: 22 September 2017

Non-coding genome

80% of the variants associated with diseases (by GWAS) are localized outside of protein-coding genes (Manolio *et al.*, 2009; Hindorff *et al.*, 2009).

>60% of the human genome is transcribed into RNAs with only 2% corresponding to proteins (Human ENCODE Consortium; Djebali *et al.*, 2012, Mouse ENCODE Consortium, 2015).

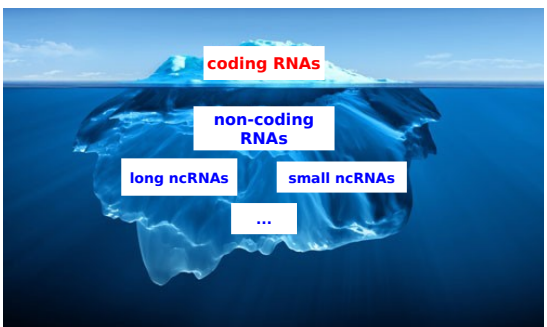
Need to identify ncRNAs to better annotate genome

Ease the interpretation of genotype to phenotype relationships.

Non-coding RNA

Different type of non-coding RNAs.
From GENCODE v27 annotation (Harrow et al., 2012):

- Total No of Genes: 58,288;
- Protein-coding genes: 19,836 (34%);
- Long non-coding RNA genes: 15,778 (27%);
- Small non-coding RNA genes: 7,569 (13%);
- ...



Long non-coding RNA characteristics

Definition:

- Transcripts without coding potential, longer than 200 nt, polyA+/- (Derrien *et al.*, 2012).

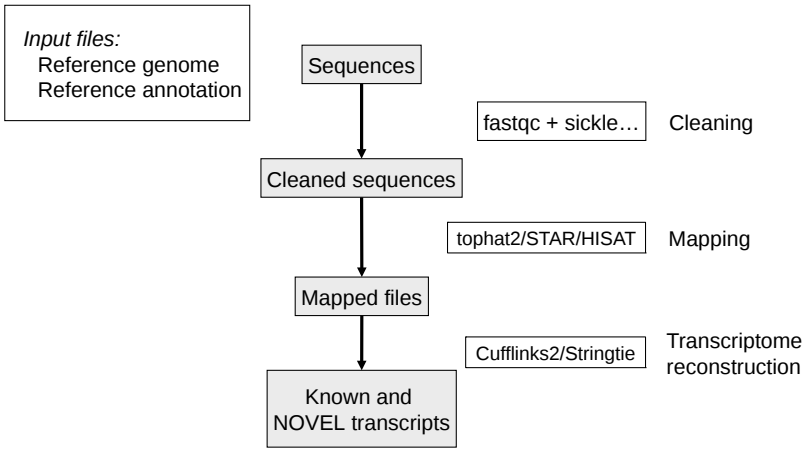
Functions (non-exhaustive):

- Can enhance or repress transcription of targeted mRNA(s);
- Can act in *cis* or in *trans*;
- Sponge for microRNAs;
- Make lncRNA - protein complexes.

Examples:

- Xist: binds to PRC2 (DNMT3A) $\rightarrow$ lncRNA-protein complexe;
- LncRNA-protein complexe $\rightarrow$ DNA hypermethylation;
- DNA hypermethylation $\rightarrow$ silencing X chromosome.

Standard pipeline for RNA-seq analysis: mRNAs + lncRNAs

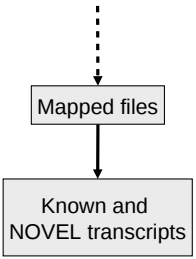


Djebali *et al.*, 2017

Bottleneck

Lots of novel transcripts.

How to deal with all assembled transcripts



Novel transcripts = lncRNAs + mRNAs + spurious transcription

Classical pipeline to annotate new transcripts:

1. Filter: remove short transcripts, smaller than 200 nt long.
2. Discriminate: determine whether the transcript is coding or not.
3. Classify: classify the lncRNAs regarding to nearest RNA genes.

Why a new tool?

Issues:

- The filter (1) and classify (3) steps are made manually;
- Only a minimal classification (3) by one of the tool, none for the others;
- No real guideline for non-model organisms.

But some tools exist to discriminate (2) between coding and non-coding RNAs.

Tools to discriminate mRNAs and lncRNAs

Alignment-based:

Advantages:

- High specificity;
- Identify conserved lncRNAs.

Drawbacks:

- Depends on the database;
- Depends on the alignment;
- Slow.

PhyloCSF (Lin, M. *et al.*, 2011), CPC (Kong, L. *et al.*, 2007), ...

Alignment-free:

Advantages:

- Usually Fast;
- Independent of alignment;
- Lineage-specific lncRNAs.

Drawbacks:

- Designed for model organisms.

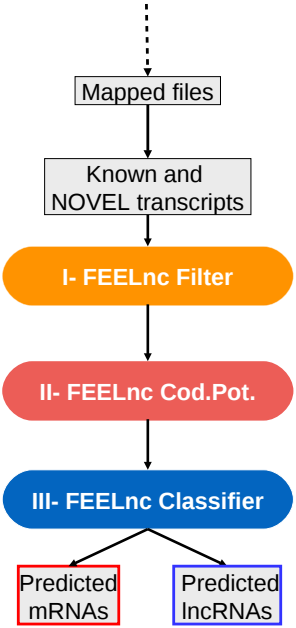
CPAT (Wang, L. *et al.* 2013), CNCI (Sun, L. *et al.* 2013), PLEK (Li, A. *et al.* 2014), ...

Our Solution

Fl**E**xible **E**xtraction of **Lnc**RNAs (FEELnc):

- All in one: formalized the filtering and classification into modules;
- Stringent set of lncRNAs and mRNAs;
- Classification regarding all RNAs, useful to get potential functional relations;
- LncRNAs detection is genome reference-free, i.e non-model species;
- Available for non-model organisms by replacing lncRNAs.

FEELnc: FlExible Extraction of LncRNAs



Transcriptome reconstruction

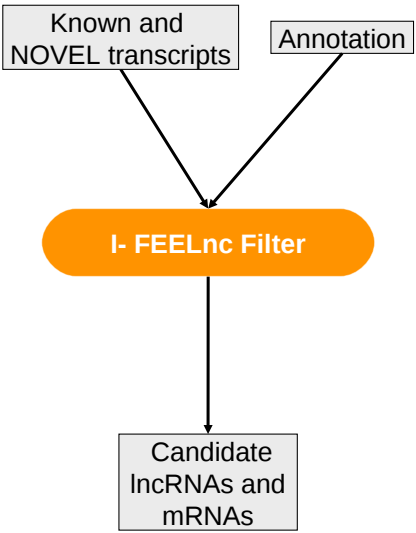
FEELnc, three independent modules:

Get all lncRNA-like transcripts.

Use a Random Forest to discriminate between mRNAs and lncRNAs.

Classify the lncRNAs regarding to nearest RNA genes (lincRNAs, antisense, host lncRNA, ...).

Filter



Aim:

- Filtering out non lncRNA-like.

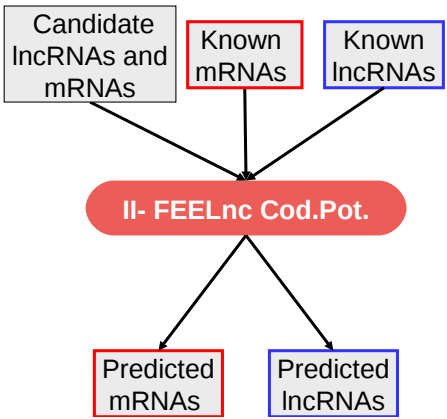
Methods:

- Remove short transcripts (< 200 nt);
- Flag transcripts overlapping known mRNAs;
- Keep or discard monoexonic transcripts, antisense or intergenic;

Next step:

- Defined coding and non-coding transcripts.

Coding Potential



Aim:

- Defines a protein-coding score and then a cutoff to differentiate mRNAs from lncRNAs.

Methods:

- Use features and machine learning, a Random Forest.

Features for the Random Forest

1. RNA size (Cabili *et al.*, 2011; Derrien *et al.*, 2012)
   (high value $\rightarrow$ mRNA)

Features for the Random Forest

1. RNA size (Cabili *et al.*, 2011; Derrien *et al.*, 2012)
   (high value $\rightarrow$ mRNA)

2. ORF coverage (ORF defined with respect to 5 modes):
   - Strict: requires start and stop;
   - Moderates: requires start or stop;
   - Relaxed: total RNA sequence.

   (high value $\rightarrow$ mRNA)

Features for the Random Forest

1. RNA size (Cabili *et al.*, 2011; Derrien *et al.*, 2012)
   (high value → mRNA)

2. ORF coverage (ORF defined with respect to 5 modes):
   - Strict: requires start and stop;
   - Moderates: requires start or stop;
   - Relaxed: total RNA sequence.

   (high value → mRNA)

3. *k*-mer scores on ORF for multiple *k*-mer sizes:
   - For a specific *k*-mer, the ratio between mRNA frequency and lncRNA frequency;
   - Collaboration with INRIA/Genscale team in Rennes (Fr), KmerInShort developed by Guillaume Rizk from GATB tools (Drezen *et al.*);
   - Very fast and parallel extraction of *k*-mer profiles up to *15*-mer.

   (high value → mRNA)

k-mer score calculation

Get the k-mer profile for a specific size k:

- For all K (e.g. TGC) of size k (e.g. 3);
- Get mRNA F_K^m and lncRNA F_K^{lnc} observed frequencies;
- Calculate a score for each K:

$$S_K^k = \frac{F_K^m}{F_K^m + F_K^{lnc}}$$

k-mer score calculation

Get the k-mer profile for a specific size k:

- For all K (e.g. TGC) of size k (e.g. 3);
- Get mRNA F_K^m and lncRNA F_K^{lnc} observed frequencies;
- Calculate a score for each K:

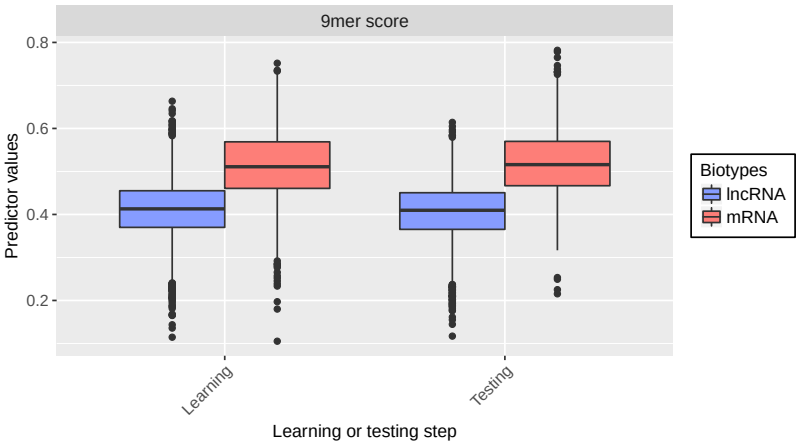
$$S_K^k = \frac{F_K^m}{F_K^m + F_K^{lnc}}$$

Get the k-mer score for a sequence X:

- For each ORF X;
- Get occurrences N_K^X of all K of size k;
- Calculate a score for the size k using all K:

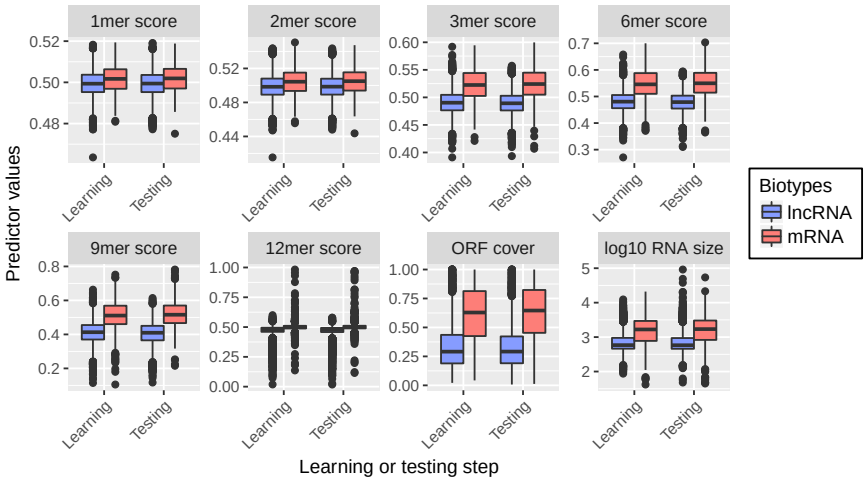
$$V_X^k = \frac{\sum_{K=1}^{4^k} S_K^k \times N_K^X}{\sum_{j=1}^{4^k} N_j^X}$$

Features: Illustration



Features comparison between 5,000 lncRNA sequences and 5,000 mRNA sequences (GENCODE v24) for the learning and the testing.

Features: Illustration



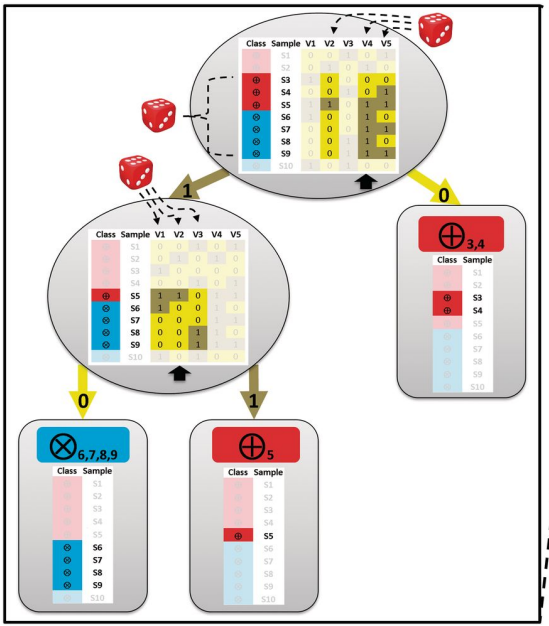
Use these features to make a Random Forest based model.

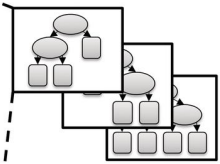
Random Forest

The **Random Forest** is a machine learning method:

- **Forest**:
  - A set of decision trees.
- **Random**:
  - Each tree is done on a sampling of the data;
  - Each node of each tree is done on a subset of the features.
- The model:
  - The **forest** of the trees made on a **sample** of the data.
- The prediction:
  - Each input sequences go through each tree;
  - Each tree vote for a sequence to be coding or non-coding;
  - Each input sequence got a score representing the number of trees which vote for the sequence to be coding.

Random Forest: Illustration



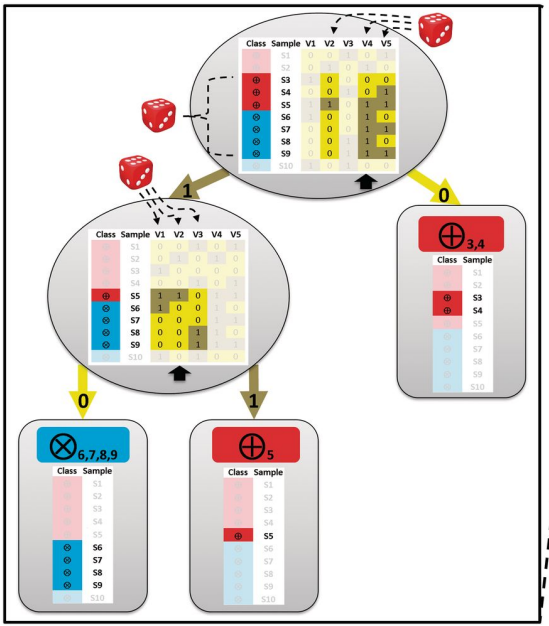


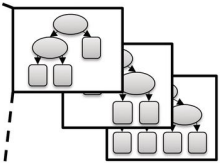
Data:

- Rows: transcripts;
- Colors: mRNAs; lncRNAs;
- Columns: features;

From Touw *et al.*, 2013

Random Forest: Illustration



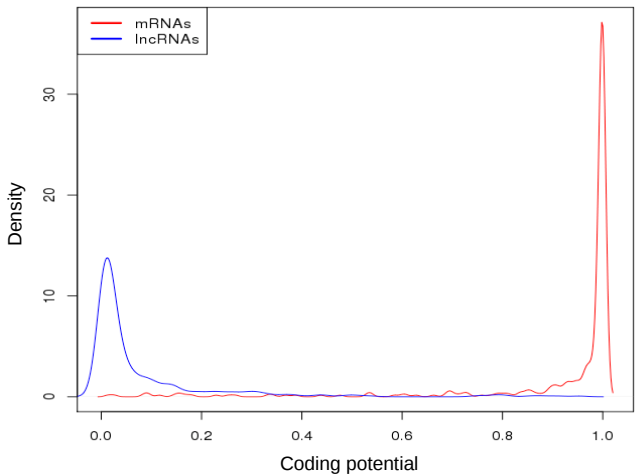


Method:

- Each tree is made with a sampling of rows;
- Each tree node is made with a feature subset;
- Chosen feature: the one which leads to the higher node purity.

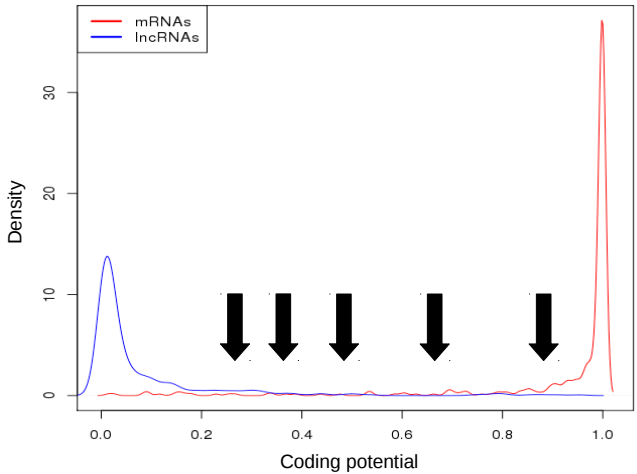
From Touw *et al.*, 2013

Coding potential score distribution



Get a coding potential score for all input sequences, with for the best case: lncRNA scores around 0 (blue) and mRNA scores around 1 (red).

Coding potential score distribution

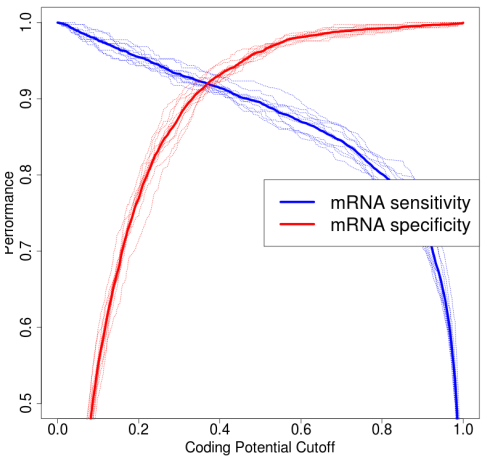


Issue

Which coding potential cutoff?

Automatic cutoff

Variation of sensitivity/specificity with the cutoff on learning dataset.

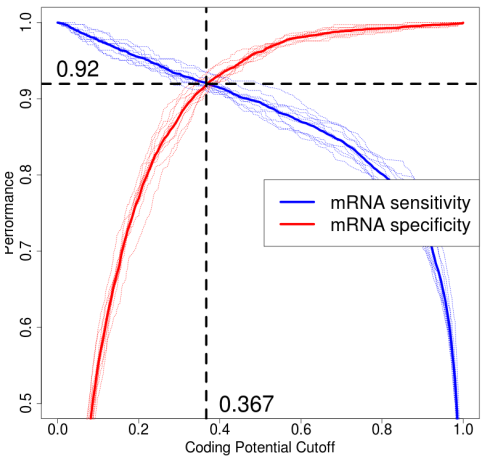


Make a 10-fold cross-validation:

- Compute performance on subset of the learning dataset.

Automatic cutoff

Variation of sensitivity/specificity with the cutoff on learning dataset.

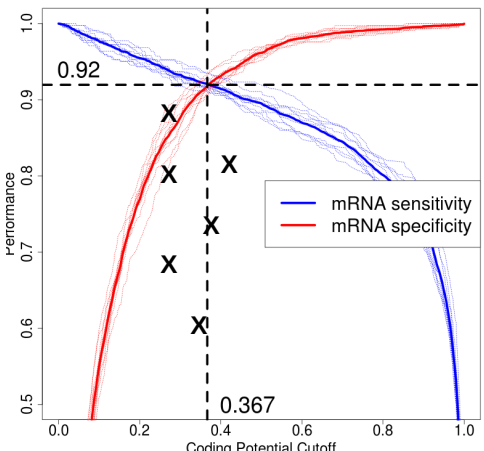


Make a 10-fold cross-validation:

- Compute performance on subset of the learning dataset;
- Use to define an optimal cutoff (0.367);
- Automatically defined as Sensitivity = Specificity (0.92).

Automatic cutoff

Variation of sensitivity/specificity with the cutoff on learning dataset.



Make a 10-fold cross-validation:

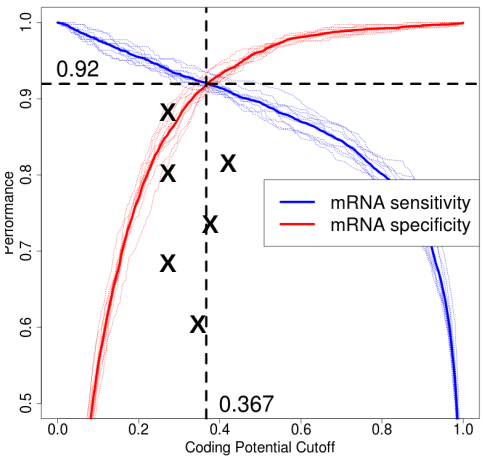
- Compute performance on subset of the learning dataset;
- Use to define an optimal cutoff (0.367);
- Automatically defined as Sensitivity = Specificity (0.92).

Issue:

- Transcripts around the cutoff, not a high confidence in the prediction.

Automatic cutoff

Variation of sensitivity/specificity with the cutoff on learning dataset.



"The (CPS) threshold is (...) somewhat arbitrary, and transcripts that reside in questionable regions of the distribution should be annotated as transcripts of unknown coding potential (TUCPs)"

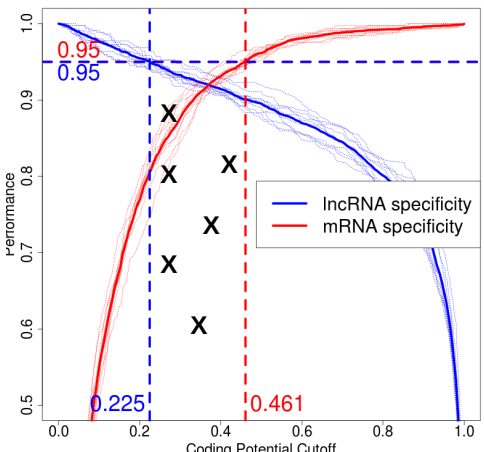
J.S. Mattick & J.L. Rinn, 2015.

Implemented FEELnc solution

Defined 2 cutoffs based on specificity.

User two cutoffs

Variation of sensitivity/specificity with the cutoff on learning dataset.

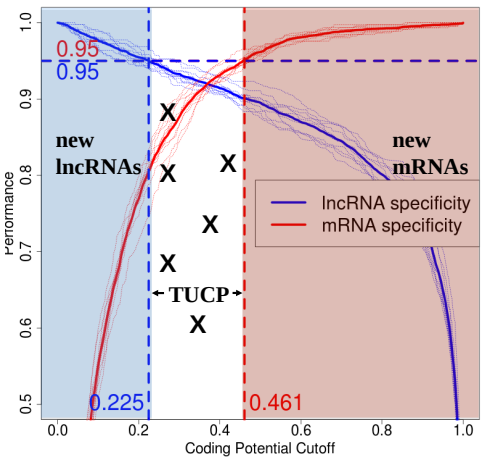


Implemented FEELnc solution:

- A user defined mRNAs and lncRNAs specificity (e.g. 0.95,0.95);
- Automatically set two cutoffs, one for mRNAs and one for lncRNAs (e.g. 0.225, 0.461).

User two cutoffs

Variation of sensitivity/specificity with the cutoff on learning dataset.



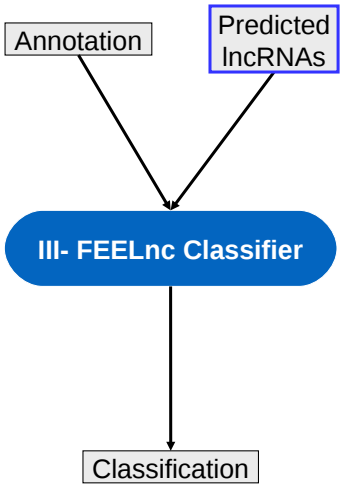
Implemented FEELnc solution:

- A user defined mRNAs and lncRNAs specificity (e.g. 0.95,0.95);
- Automatically set two cutoffs, one for mRNAs and one for lncRNAs (e.g. 0.225, 0.461).

With two cutoffs, definition of a new class:

- Transcript of Unknown Coding Potential (TUCP).

Classifier



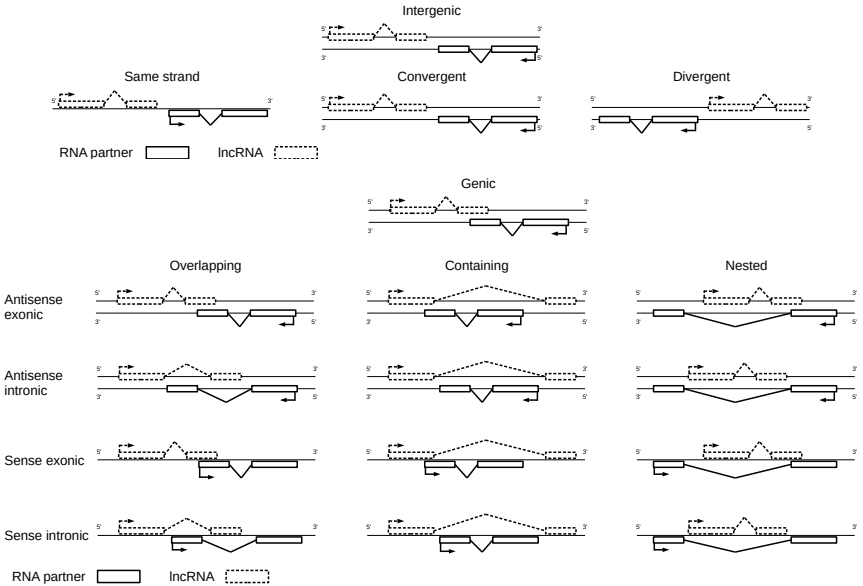
Aim:

- Predict potential functional relationships between lncRNA transcripts and RNA transcripts.

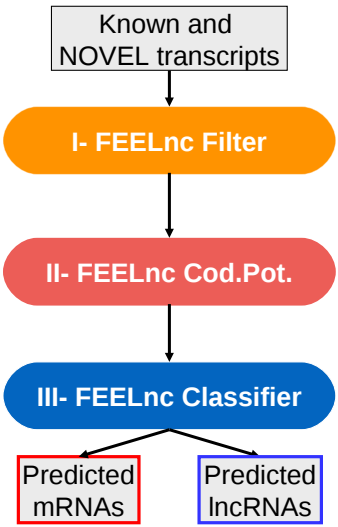
Method:

- Formalized sub-classes of genomic classification genic and intergenic;
- Get direction of the relation;
- Use a sliding window around lncRNAs;
- Get the relations for all RNA inside the window.

Classes


Intergenic
Same strand
Convergent
Divergent
RNA partner
lncRNA
Genic
Overlapping
Containing
Nested
Antisense exonic
Antisense intronic
Sense exonic
Sense intronic
RNA partner
lncRNA

FEELnc: FlExible Extraction of LncRNAs



FEELnc, three independent modules:

Get all lncRNA-like transcripts.

Use a Random Forest to discriminate between mRNAs and lncRNAs.

Classify the lncRNAs regarding the nearest transcript.

Need to compare FEELnc predictions with state of the art methods.

Benchmarking the Coding Potential module

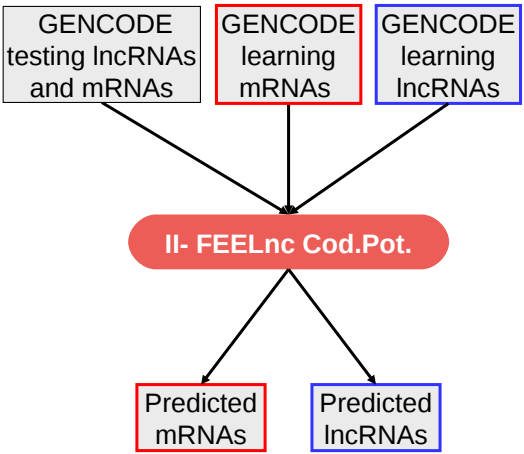
Compare the FEELnc Coding Potential module against 5 methods:

- CPAT (Wang *et al.*, 2013);
- CNCI (Sun *et al.*, 2013);
- PLEK (Li *et al.*, 2014);
- CPC (Kong *et al.*, 2007);
- PhyloCSF (Lin *et al.*, 2011).

Use 5 performance measures to compare methods:

- Sensitivity;
- Specificity;
- Precision;
- Accuracy;
- Matthews Correlation Coefficient (MCC): **summarizes others**
  1: good predictions; 0: random predictions; -1: opposed predictions.

Benchmarking



Data:

- Human GENCODE (v24) training and testing data;
- Only one transcript per gene was extracted, no common genes for learning and testing;
- 5,000 mRNAs and 5,000 lncRNAs on training and testing datasets.

Methods used with default parameters.

Benchmarking results

program	sensitivity	specificity	precision	accuracy	MCC
FEELnc	**0.923**	0.915	0.916	**0.919**	**0.838**
CPAT	0.899	0.924	0.922	0.912	0.823
CNCI	0.829	0.979	0.975	0.904	0.817
PLEK	0.732	**0.985**	**0.981**	0.858	0.741
PhyloCSF*	0.906	0.802	0.820	0.854	0.712
CPC*	0.699	0.739	0.728	0.719	0.438

Best score

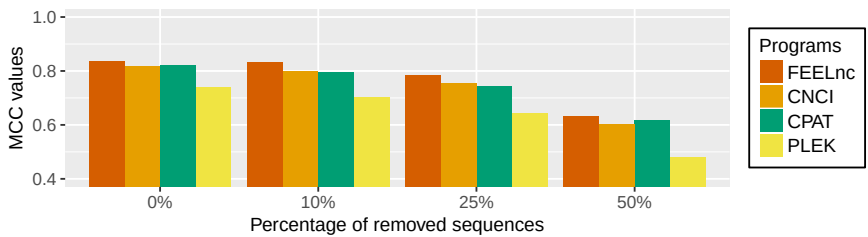
*: alignment-based

- FEELnc performs similarly or better;
- On this benchmark, alignment-free better than alignment-based.

Incomplete transcripts

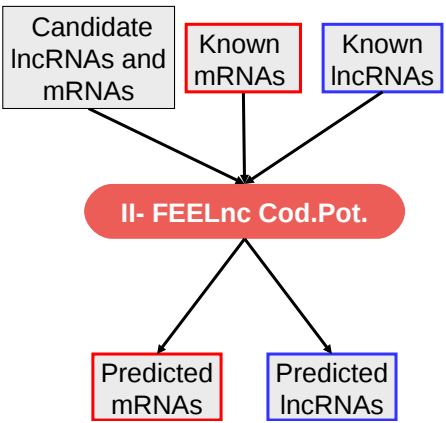
RNA-seq or model reconstruction can generate truncated/incomplete transcript models (Steijger *et al.*, 2013).

Modified the benchmark dataset by removing percentage of transcripts, either in 5' or 3'.



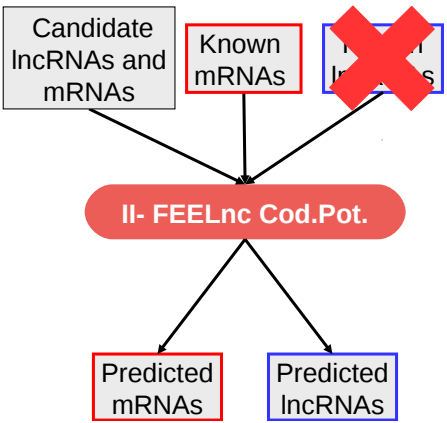
Good performance regarding other methods even with modified datasets.

What about non-model species?



Classification tools for lncRNAs work well on species with known lncRNAs.

What about non-model species?



Classification tools for lncRNAs work well on species with known lncRNAs.

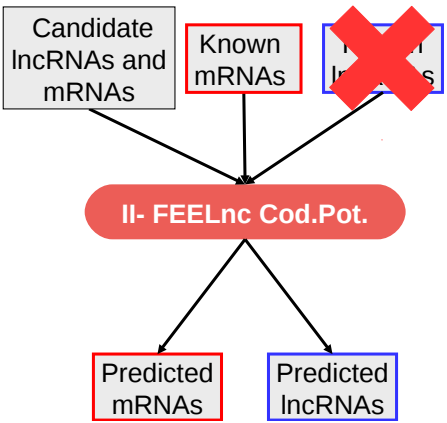
Issues

What if no lncRNAs are available, i.e. non model species?

Some solutions

1. Mimic lncRNAs with other sequences for the learning step;
2. Use lncRNA sequences from evolutionary related species.

What about non-model species?



Classification tools for lncRNAs work well on species with known lncRNAs.

Issues

What if no lncRNAs are available, i.e. non model species?

Some solutions

1. Mimic lncRNAs with other sequences for the learning step;
2. Use lncRNA sequences from evolutionary related species.

Mimic lncRNA sequences

How to mimic lncRNA sequences?

1. LncRNAs are non-coding $\rightarrow$ extract non-coding sequences;
2. LncRNAs can result from the pseudogenization of protein coding genes (Duret *et al.*, 2006) $\rightarrow$ modify mRNA sequences.

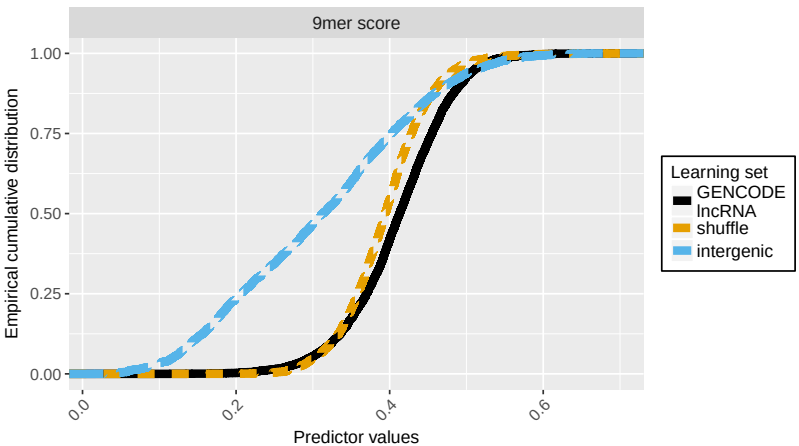
FEELnc methods to mimic lncRNA sequences:

1. Intergenic module: randomly extract genomic intergenic sequences;
2. Shuffle module: shuffle mRNA learning sequences while preserving the *7*-mer frequencies using Ushuffle (Jiang *et al.*, 2008).

Mimic lncRNA sequences

How to mimic lncRNA sequences?

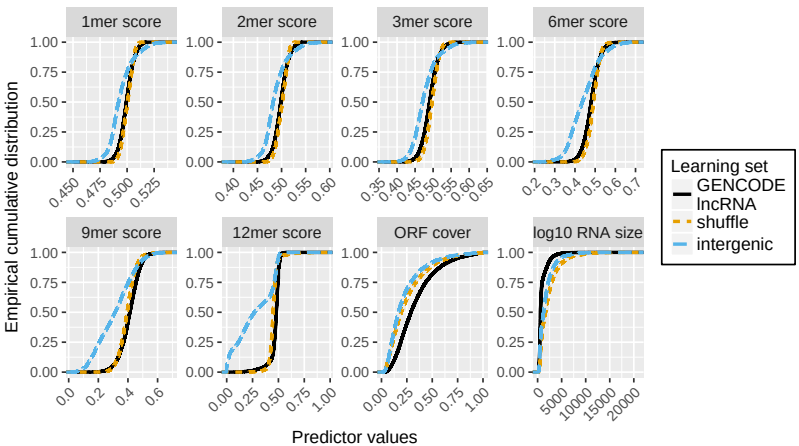
1. LncRNAs are non-coding → extract non-coding sequences;
2. LncRNAs can result from the pseudogenization of protein coding genes (Duret *et al.*, 2006) → modify mRNA sequences.



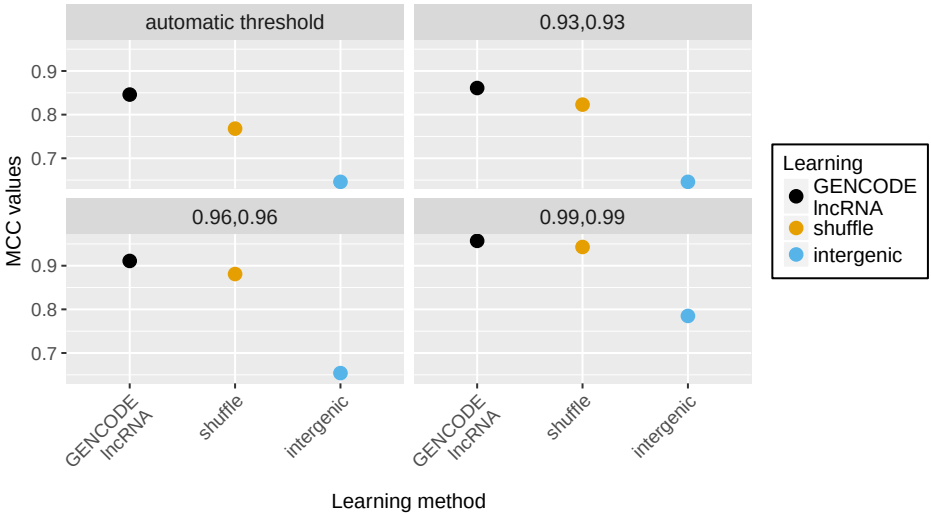
Mimic lncRNA sequences

How to mimic lncRNA sequences?

1. LncRNAs are non-coding → extract non-coding sequences;
2. LncRNAs can result from the pseudogenization of protein coding genes (Duret *et al.*, 2006) → modify mRNA sequences.

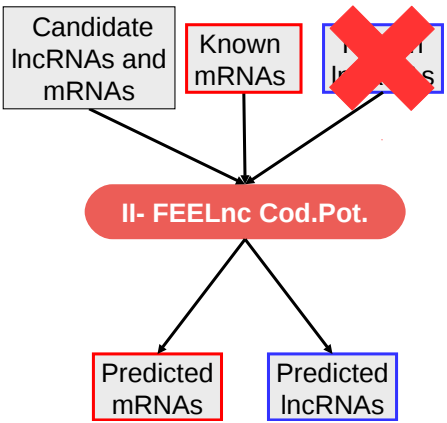


Mimic lncRNA sequences: Results



Performance on the human testing dataset using as learning lncRNAs either the GENCODE lncRNAs, the shuffle module or the intergenic module.

What about non-model species?



Classification tools for lncRNAs work well on species with known lncRNAs.

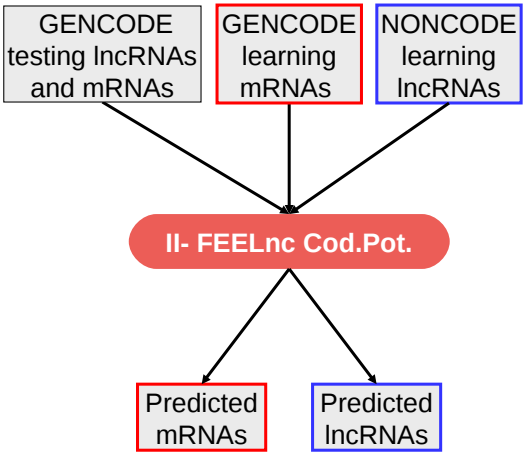
Issues

What if no lncRNAs are available, i.e. non model species?

Some solutions

1. Mimic lncRNAs with other sequences for the learning step;
2. Use lncRNA sequences from evolutionary related species.

Evolutionary related lncRNA sequences



Train the Random Forest:

- Use mRNAs of the species;
- Use lncRNAs from the evolutionary related species.

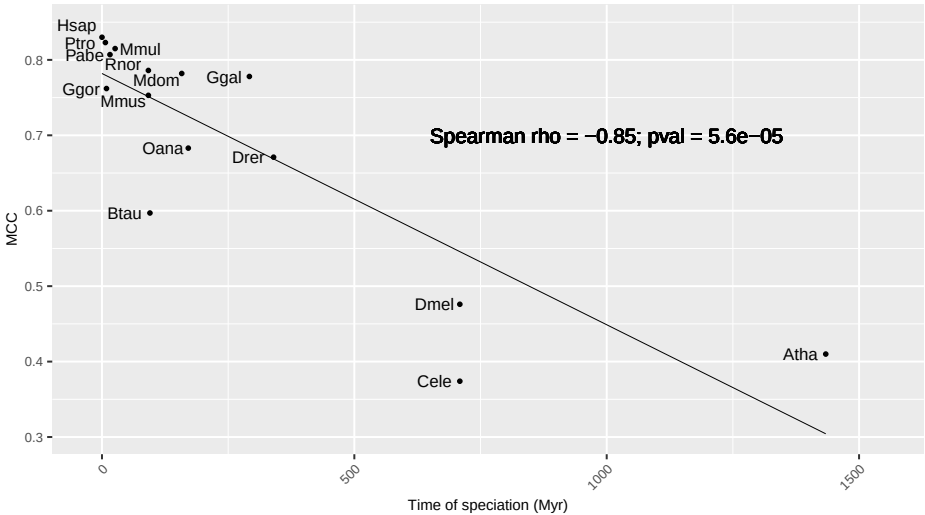
Apply the model:

- On transcript models of the species.

How to test?

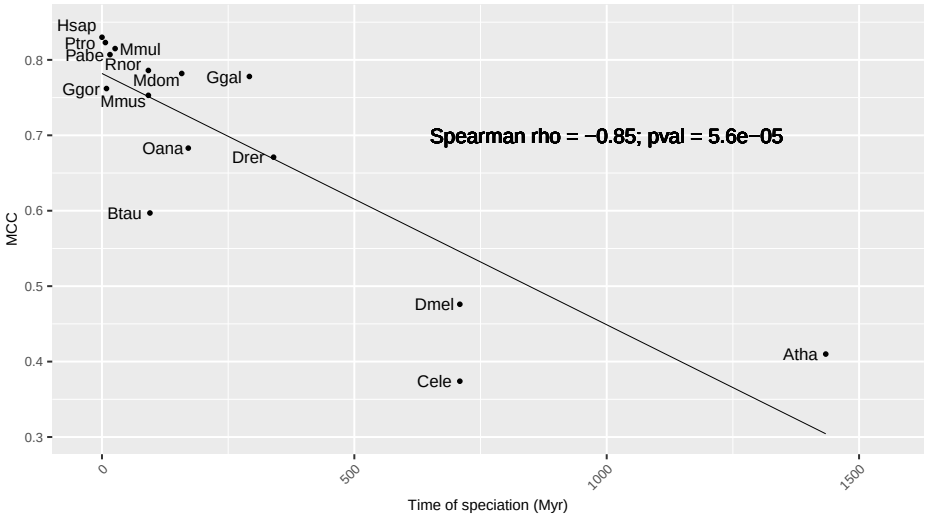
Learning lncRNAs sequences from the NONCODE 2016 database (Zhao *et al.*, 2016)

Evolutionary related lncRNA sequences: Results



FEELnc performance and times of speciation are anti-correlated.

Evolutionary related lncRNA sequences: Results

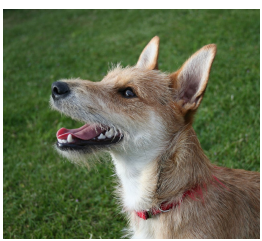


Shuffle module: 0.748 MCC → ~100 Myr.

Application: Dog

Dog (Wucher *et al.*, 2017):

- Collaboration with the European LUPA Consortium (Michel Georges) and the BROAD institute (Kerstin Lindblad-Toh);
- 16 tissues;
- 20 RNA-seq;
- ~2,500 new lncRNA genes;
- ~10,000 new lncRNA transcripts.



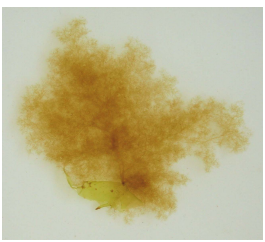
Dog (Chris Barber, Wikipedia)

Applications: Others

Chicken (Andrei Niemimäki, Wikipedia)

Chicken (Muret *et al.*, 2017):

- With Sandrine Lagarrigue (INRA, Agrocampus, FR);
- Adipose and liver tissues;
- 16 RNA-seq;
- ~2,200 new lncRNA genes.



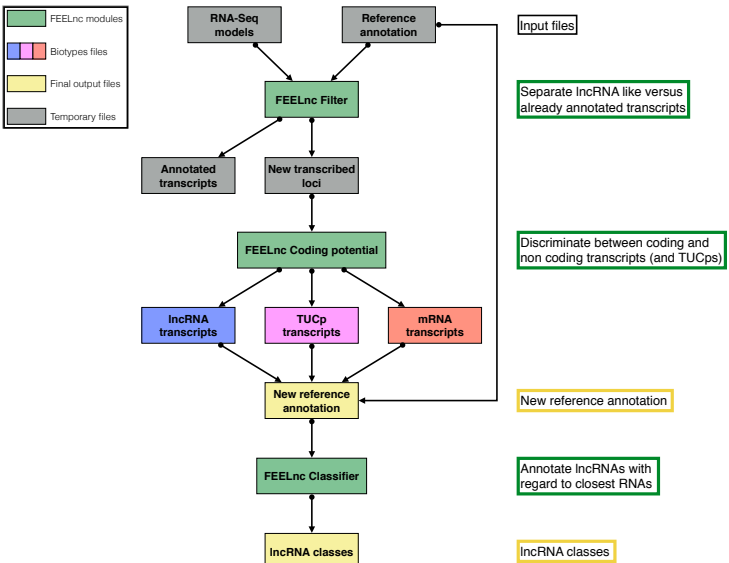
Ectocarpus (Akirapeters, Wikipedia)

Ectocarpus (algae) (Cormier *et al.*, 2016):

- With Mark Cock (CNRS,Roscoff, FR);
- Male and femelle gametophytes;
- 10 RNA-seq;
- ~700 new lncRNA genes;
- First lncRNA catalogue in algae.

One tool: three applications

FEELnc: from transcript models to lncRNA classifications.



Conclusion

User friendly:

- Automatic threshold;
- Easy to use.

Flexible:

- Set two specificity thresholds for stringent predictions;
- Five ORF type definitions;
- Coding potential can be used on FASTA or GTF.

Non-model species compatible:

- Mimic lncRNAs by shuffling mRNA sequences;
- Coding potential module is alignment- and genome reference-free;
- Guideline for species without annotated lncRNAs.

Performs similar or better than other tools.

Availability

FEELnc on github:

https://github.com/tderrien/FEELnc

A nextflow/docker implementation of STAR/Cufflinks/FEELnc pipeline has been developed by Evan FLODEN:

https://github.com/skptic/lncRNA-Annotation-nf

Published in Nucleic Acids Research, along with a dog extended annotation: Wucher *et al.*, 2017.

All data (included benchmarking data), command lines and scripts to make figures are available through Supplementary:

http://nar.oxfordjournals.org/content/early/2017/01/03/nar.gkw1306/suppl/DC1

Prospects

Method:

- Add new features, e.g transcript expressions or exons number;
- Modified the 12-mer score, e.g. translate amino acids;
- Predict pseudogenes (can be still reference-free?).

Bioanalyses:

- FEELnc used in the FAANG consortium, i.e. farming animals;
- Improve lncRNA classification by adding new data, as chromosome configuration capture (Hi-C);
- Use and integrate lncRNA classification with multi-omics data.

Acknowledgments

CNRS - IGDR - Dog genetics team (Rennes):
Thomas Derrien, **Christophe Hitte**, Catherine André, Céline Le Béguec

INRA/INRIA - IGEPP/IRISA - EGI/Genscale team (Rennes):
Fabrice Legeai, **Guillaume Rizk**

INRA - PEGASE (Rennes):
Kévin Muret, Sandrine Lagarrigue

CNRS - Algal Genetics Group (Roscoff):
Alexandre Cormier, Erwan Corre, Susana M. Coelho, Mark Cock

INRA - Toulouse:
Sarah Djebali, Sylvain Foissac

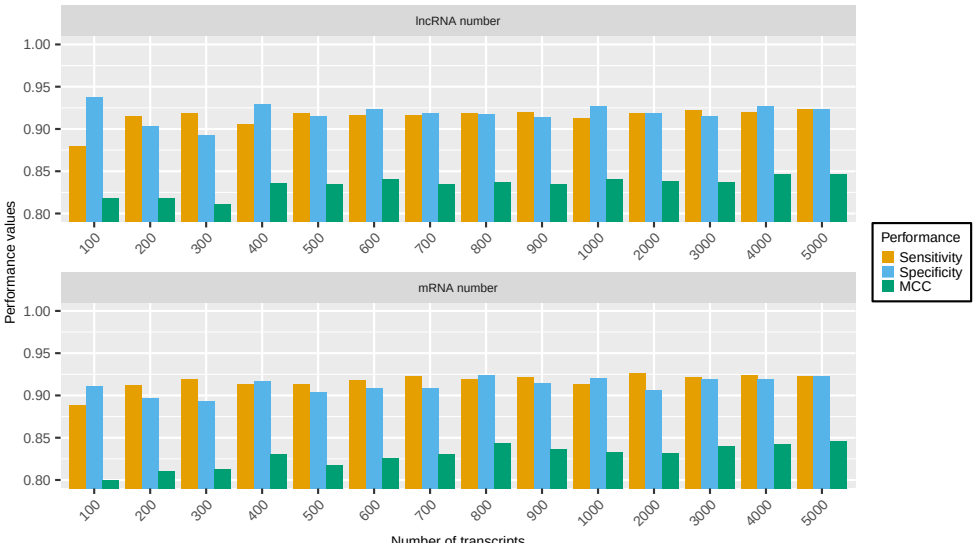
CRG - Comparative Bioinformatics (Barcelona):
Evan Floden

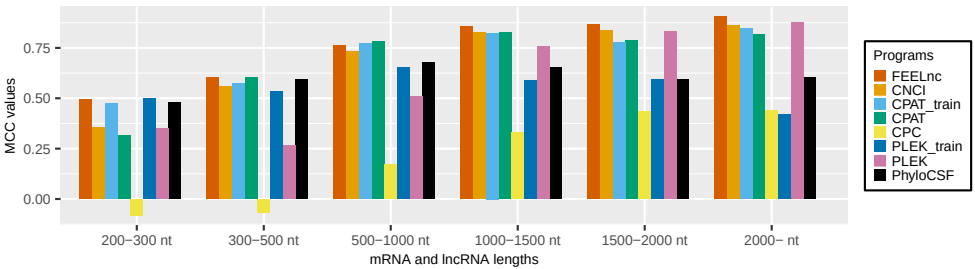
CRG - Computational Biology of RNA Processing (Barcelona):
Roderic Guigó and all lab's members.

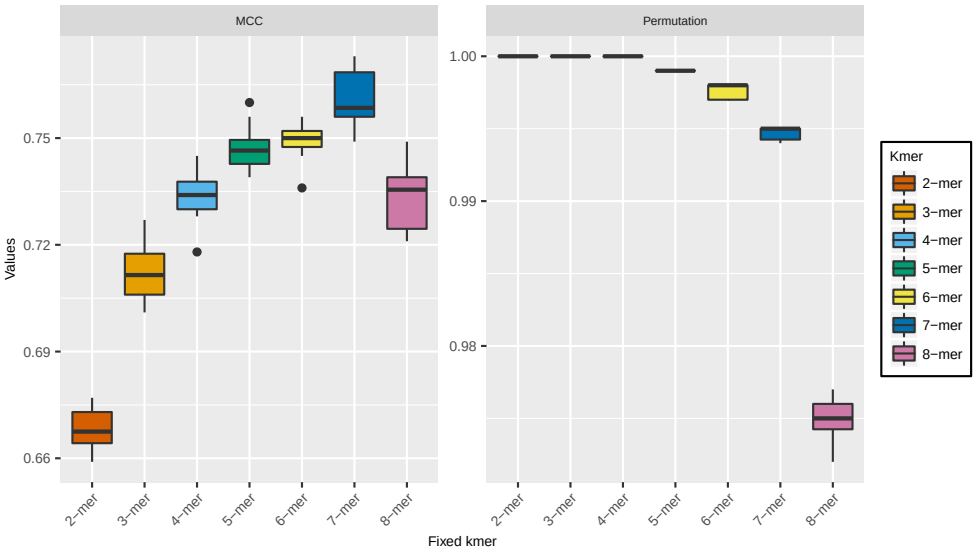
Nathalie Villa-Vialaneix

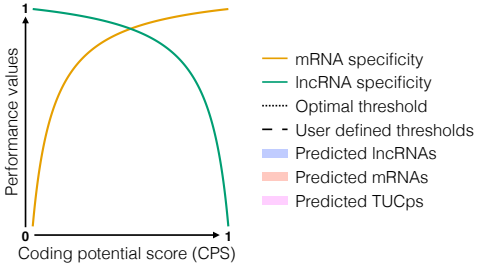
And all of you for your attention!

Annexes

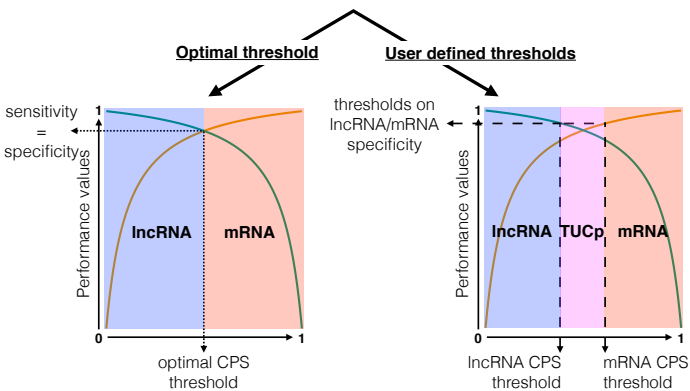

lncRNA number
mRNA number
Performance values
1.00
0.95
0.90
0.85
0.80
100
200
300
400
500
600
700
800
900
1000
2000
3000
4000
5000
Number of transcripts
Performance
Sensitivity
Specificity
MCC

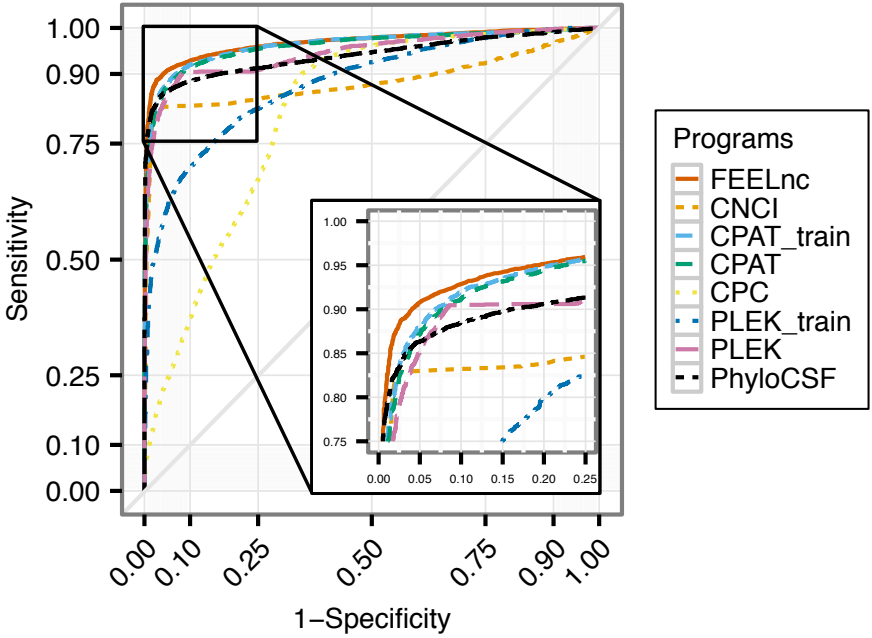

MCC values
0.75
0.50
0.25
0.00
200−300 nt
300−500 nt
500−1000 nt
1000−1500 nt
1500−2000 nt
2000− nt
mRNA and lncRNA lengths
Programs
FEELnc
CNCI
CPAT_train
CPAT
CPC
PLEK_train
PLEK
PhyloCSF


MCC
Permutation
Values
0.75
0.72
0.69
0.66
1.00
0.99
0.98
2−mer
3−mer
4−mer
5−mer
6−mer
7−mer
8−mer
Fixed kmer
Kmer
2−mer
3−mer
4−mer
5−mer
6−mer
7−mer
8−mer



Two-graph ROC curves on 10-fold cross-validation




Sensitivity
1.00
0.90
0.75
0.50
0.25
0.10
0.00
1−Specificity
0.00
0.10
0.25
0.50
0.75
0.90
1.00
0.95
0.85
0.80
0.05
0.15
0.20
Programs
FEELnc
CNCI
CPAT_train
CPAT
CPC
PLEK_train
PLEK
PhyloCSF

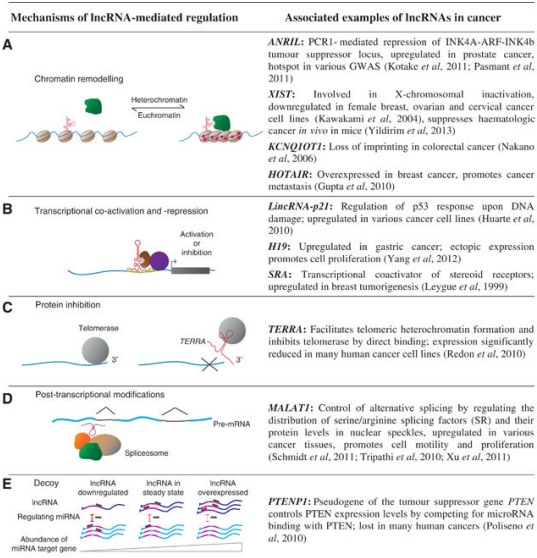
Original sequence:
AGACTTAGCA
Original count:

AC	AG	CA	CT	TA	TT	GA	GC
1	2	1	1	1	1	1	1

Ushuffle with k-mer size $= 2$

Permuted sequence:
ACTTAGCAGA
Permuted count:

AC	AG	CA	CT	TA	TT	GA	GC
1	2	1	1	1	1	1	1



From: Cheetham, S. W., Gruhl, F., Mattick, J. S., & Dinger, M. E. (2013). Long noncoding RNAs and the genetics of cancer. British journal of cancer, 108(12), 2419.