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

FEELnc: FIExible Extraction of LncRNAs

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

Séminaire MIAT: 22 September 2017
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 types 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 IncRNA - protein complexes.

Examples:
- Xist: binds to PRC2 (DNMT3A) → IncRNA-protein complexe;
- LncRNA-protein complexe → DNA hypermethylation;
- DNA hypermethylation → silencing X chromosome.
Standard pipeline for RNA-seq analysis: mRNAs + lncRNAs

**Input files:**
- Reference genome
- Reference annotation

**Processes:**
- Sequences
  - Cleaning: fastqc + sickle...
  - Mapping: tophat2/STAR/HISAT
  - Transcriptome reconstruction: Cufflinks2/Stringtie

**Outputs:**
- Cleaned sequences
- Mapped files
- Known and NOVEL transcripts

Djebali *et al.*, 2017

**Bottleneck**
Lots of novel transcripts.
How to deal with all assembled transcripts

Novel transcripts = IncRNAs + 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 IncRNAs 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 IncRNAs

Alignment-based:
Advantages:
- High specificity;
- Identify conserved IncRNAs.

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 IncRNAs.

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**

**FiE**xible **Ex**traction of **Lnc**RNAs (FEELnc):
- All in one: formalized the filtering and classification into modules;
- Stringent set of IncRNAs 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 IncRNAs.
FEELnc, three independent modules:

I- FEELnc Filter
Get all IncRNA-like transcripts.

II- FEELnc Cod.Pot.
Use a Random Forest to discriminate between mRNAs and IncRNAs.

III- FEELnc Classifier
Classify the IncRNAs regarding to nearest RNA genes (lincRNAs, antisense, host IncRNA, ...).

Transcriptome reconstruction
Filter

Known and NOVEL transcripts

Annotation

I- FEELnc Filter

Candidate IncRNAs and mRNAs

Aim:
- Filtering out non IncRNA-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 IncRNAs.

**Methods:**
- Use features and machine learning, a Random Forest.

**Diagram:**
- Candidate IncRNAs and mRNAs
- Known mRNAs
- Known IncRNAs
- Predicted mRNAs
- Predicted IncRNAs
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 → 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)
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)

3. $k$-mer scores on ORF for multiple $k$-mer sizes:
   - For a specific $k$-mer, the ratio between mRNA frequency and IncRNA 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 $\rightarrow$ mRNA)
Get the $k$-mer profile for a specific size $k$:

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

\[ S^k_K = \frac{F^m_K}{F^m_K + F^{lnc}_K} \]
**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 $F^m_K$ and lncRNA $F^{\text{Inc}}_K$ observed frequencies;
- Calculate a score for each $K$:

$$S^k_K = \frac{F^m_K}{F^m_K + F^{\text{Inc}}_K}$$

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 $K$:

$$V^k_X = \frac{\sum_{K=1}^{4^k} S^k_K \times N^K_X}{\sum_{j=1}^{4^k} N^j_X}$$
Features comparison between 5,000 lncRNA sequences and 5,000 mRNA sequences (GENCODE v24) for the learning and the testing.
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; IncRNAs;
- Columns: features;

From Touw et al., 2013
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
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.
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)"

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 IncRNAs specificity (e.g. 0.95, 0.95);
- Automatically set two cutoffs, one for mRNAs and one for IncRNAs (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).
**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

- **Same strand**
  - 5' 3'
- **Intergenic**
  - 5' 3'
  - 3' 5'
- **Convergent**
  - 5' 3'
  - 3' 5'
- **Divergent**
  - 5' 3'
  - 3' 5'

**RNA partner**

**lncRNA**

**Genic**

**Overlapping**

**Containing**

**Nested**

**Antisense exonic**

**Antisense intronic**

**Sense exonic**

**Sense intronic**

**RNA partner**

**lncRNA**

23
FEELnc: FlExible Extraction of LncRNAs

FEELnc, three independent modules:

I- FEELnc Filter

Get all lncRNA-like transcripts.

II- FEELnc Cod.Pot.

Use a Random Forest to discriminate between mRNAs and lncRNAs.

III- FEELnc Classifier

Classify the lncRNAs regarding the nearest transcript.

Predicted mRNAs

Predicted lncRNAs

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**

**GENCODE testing IncRNAs and mRNAs**

**GENCODE learning mRNAs**

**GENCODE learning IncRNAs**

---

**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 IncRNAs on training and testing datasets.

Methods used with default parameters.
**Benchmarking results**

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

**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 IncRNAs work well on species with known IncRNAs.
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 IncRNAs work well on species with known IncRNAs.

Issues

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

Some solutions

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

Candidate IncRNAs and mRNAs

Known mRNAs

II- FEELnc Cod.Pot.

Predicted mRNAs

Predicted IncRNAs
Mimic IncRNA sequences

How to mimic IncRNA 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 IncRNA 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 IncRNA sequences

How to mimic IncRNA 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 IncRNA sequences: Results

Performance on the human testing dataset using as learning IncRNAs either the GENCODE IncRNAs, the shuffle module or the intergenic module.
Classification tools for IncRNAs work well on species with known IncRNAs.

**Issues**

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

**Some solutions**

1. Mimic IncRNAs with other sequences for the learning step;
2. Use IncRNA sequences from evolutionary related species.
Evolutionary related IncRNA sequences

Train the Random Forest:
- Use mRNAs of the species;
- Use IncRNAs from the evolutionary related species.

Apply the model:
- On transcript models of the species.

How to test?
Learning IncRNAs sequences from the NONCODE 2016 database (Zhao et al., 2016)
Evolutionary related IncRNA sequences: Results

**FEELnc performance and times of speciation are anti-correlated.**
Evolutionary related lncRNA sequences: Results

Spearman rho = -0.85; pval = 5.6e-05

Shuffle module: 0.748 MCC $\rightarrow$ $\sim$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 IncRNA genes;
- ~10,000 new IncRNA transcripts.
Applications: Others

Chicken (Muret et al., 2017):
- With Sandrine Lagarrigue (INRA, Agrocampus, FR);
- Adipose and liver tissues;
- 16 RNA-seq;
- ~2,200 new IncRNA genes.

Ectocarpus (algae) (Cormier et al., 2016):
- With Mark Cock (CNRS, Roscoff, FR);
- Male and femelle gametophytes;
- 10 RNA-seq;
- ~700 new IncRNA genes;
- First IncRNA catalogue in algae.
One tool: three applications

FEELnc: from transcript models to IncRNA 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 IncRNAs by shuffling mRNA sequences;
- Coding potential module is alignment- and genome reference-free;
- Guideline for species without annotated IncRNAs.

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/IncRNA-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 IncRNA classification by adding new data, as chromosome configuration capture (Hi-C);
- Use and integrate IncRNA 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 -
EGL/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
Two-graph ROC curves on 10-fold cross-validation

- mRNA specificity
- IncRNA specificity
- Optimal threshold
- User defined thresholds
- User defined thresholds

Coding potential score (CPS)

Performance values

Sensitivity = specificity

Optimal threshold

User defined thresholds

Predicted IncRNAs

Predicted mRNAs

Predicted TUCps
Original sequence:
AGACTTAGCA
Original count:

<table>
<thead>
<tr>
<th>AC</th>
<th>AG</th>
<th>CA</th>
<th>CT</th>
<th>TA</th>
<th>TT</th>
<th>GA</th>
<th>GC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Ushuffle with $k$-mer size = 2

Permuted sequence:
ACTTAGCAGA
Permuted count:

<table>
<thead>
<tr>
<th>AC</th>
<th>AG</th>
<th>CA</th>
<th>CT</th>
<th>TA</th>
<th>TT</th>
<th>GA</th>
<th>GC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Introduction FEELnc description Benchmark Non-model species Conclusion Prospects

Mechanisms of IncRNA-mediated regulation

A

Chromatin remodelling

B

Transcriptional co-activation and -repression

C

Protein inhibition

D

Post-transcriptional modifications

E

Decoy

 Associated examples of IncRNAs in cancer

ANRIL: PCR1-mediated repression of INK4A-ARF-INK4b tumour suppressor loci, upregulated in prostate cancer, hotspot in various GWAS (Kotake et al., 2011; Pasquali et al., 2011)

MST1: Involved in X-chromosomal inactivation, downregulated in female breast, ovarian and cervical cancer cell lines (Kawakami et al., 2004), suppresses haematologic cancer in vivo in mice (Tüttö et al., 2013)

KCNQ1OT1: Loss of imprinting in colorectal cancer (Nakano et al., 2006)

HOTAIR: Overexpressed in breast cancer, promotes cancer metastasis (Gupta et al., 2010)

LineRNA-p21: Regulation of p53 response upon DNA damage; upregulated in various cancer cell lines (Huarte et al., 2010)

IncRNA-H19: Upregulated in gastric cancer; ectopic expression promotes cell proliferation (Yang et al., 2012)

SRA: Transcriptional coactivator of steroidal receptors; upregulated in breast tumorigenesis (Leygue et al., 1999)

TEFRA: Facilitates telomeric heterochromatin formation and inhibits telomerase by direct binding; expression significantly reduced in many human cancer cell lines (Redon et al., 2010)

MALAT1: Control of alternative splicing by regulating the distribution of serine/arginine splicing factors (SR) and their protein levels in nuclear speckles, upregulated in various cancer tissues, promotes cell motility and proliferation (Schmidt et al., 2011; Tripathi et al., 2010; Xu et al., 2011)

PTENP1: Pseudogene of the tumour suppressor gene PTEN controls PTEN expression levels by competing for microRNA binding with PTEN; lost in many human cancers (Poliseno et al., 2010)