Long non-coding RNAs (lncRNAs) annotation

**Stéphanie ROBIN, BIPAA platform, INRAE /
GenOuest platform, Rennes**

From CATI BARIC, INRAE

Non-coding RNAs importance

- 80% of the variants associated with diseases are not located on protein-coding genes. (Manolio et al, Hindorrf et al)

- More than 60% of the human genome is transcribed into RNA (75% by primary transcripts), but only 2% will be translated into proteins

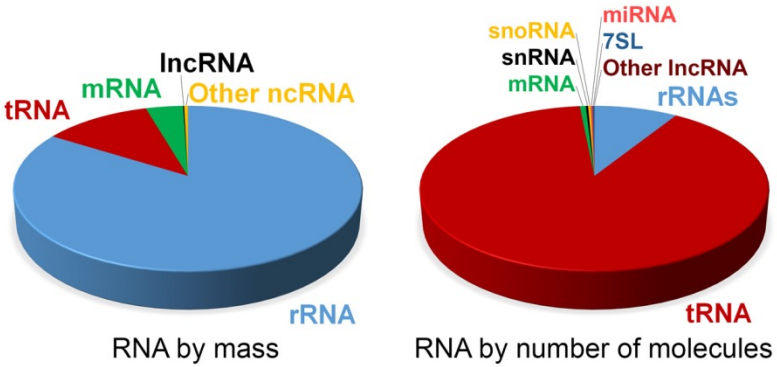
=> Need to annotate <u>non-coding RNAs</u> to increase genotype / phenotype understanding.

Impact of non-coding ?

The different types of RNAs

Type	Percent of total RNA by mass	Molecules per cell	Average size (kb)	Total weight picograms/cell	Notes	Reference
rRNAs	80 to 90	3–10 × 10^6 (ribosomes)	6.9	10 to 30		Blobel and Potter (1967), Wolf and Schlessinger (1977), Duncan and Hershey (1983)
tRNA	10 to 15	3–10 × 10^7	<0.1	1.5 to 5	About 10 tRNA molecules /ribosome	Waldron and Lacroute (1975)
mRNA	3 to 7	3–10 × 10^5	1.7	0.25 to 0.9		Hastie and Bishop (1976), Carter et al. (2005)
hnRNA (pre-mRNA)	0.06 to 0.2	1–10 × 10^3	10*	0.004 to 0.03	Estimated at 2–4% of mRNA by weight	Mortazavi et al. (2008), Menet et al. (2012)
Circular RNA	0.002 to 0.03	3–20 × 10^3	~0.5	0.0007 to 0.005	Estimated at 0.1–0.2% of mRNA**	Salzman et al. (2012), Guo et al. (2014)
snRNA	0.02 to 0.3	1–5 × 10^5	0.1–0.2	0.008 to 0.04		Kiss and Filipowicz (1992), Castle et al. (2010)
snoRNA	0.04 to 0.2	2–3 × 10^5	0.2	0.02 to 0.03		Kiss and Filipowicz (1992), Cooper (2000), Castle et al. (2010)
miRNA	0.003 to 0.02	1–3 × 10^5	0.02	0.001 to 0.003	About 10^5 molecules per 10 pg total RNA	Bissels et al. (2009)
7SL	0.01 to 0.2	3–20 × 10^4	0.3	0.005 to 0.03	About 1–2 SRP molecules/100 ribosomes	Raue et al. (2007), Castle et al. (2010)
Xist	0.0003 to 0.02	0.1–2 × 10^3	2.8	0.0001 to 0.003		Buzin et al. (1994), Castle et al. (2010)
Other lncRNA	0.03 to 0.2	3–50 × 10^3	1	0.002 to 0.03	Estimated at 1–4% of mRNA by weight	Mortaza et al. (2

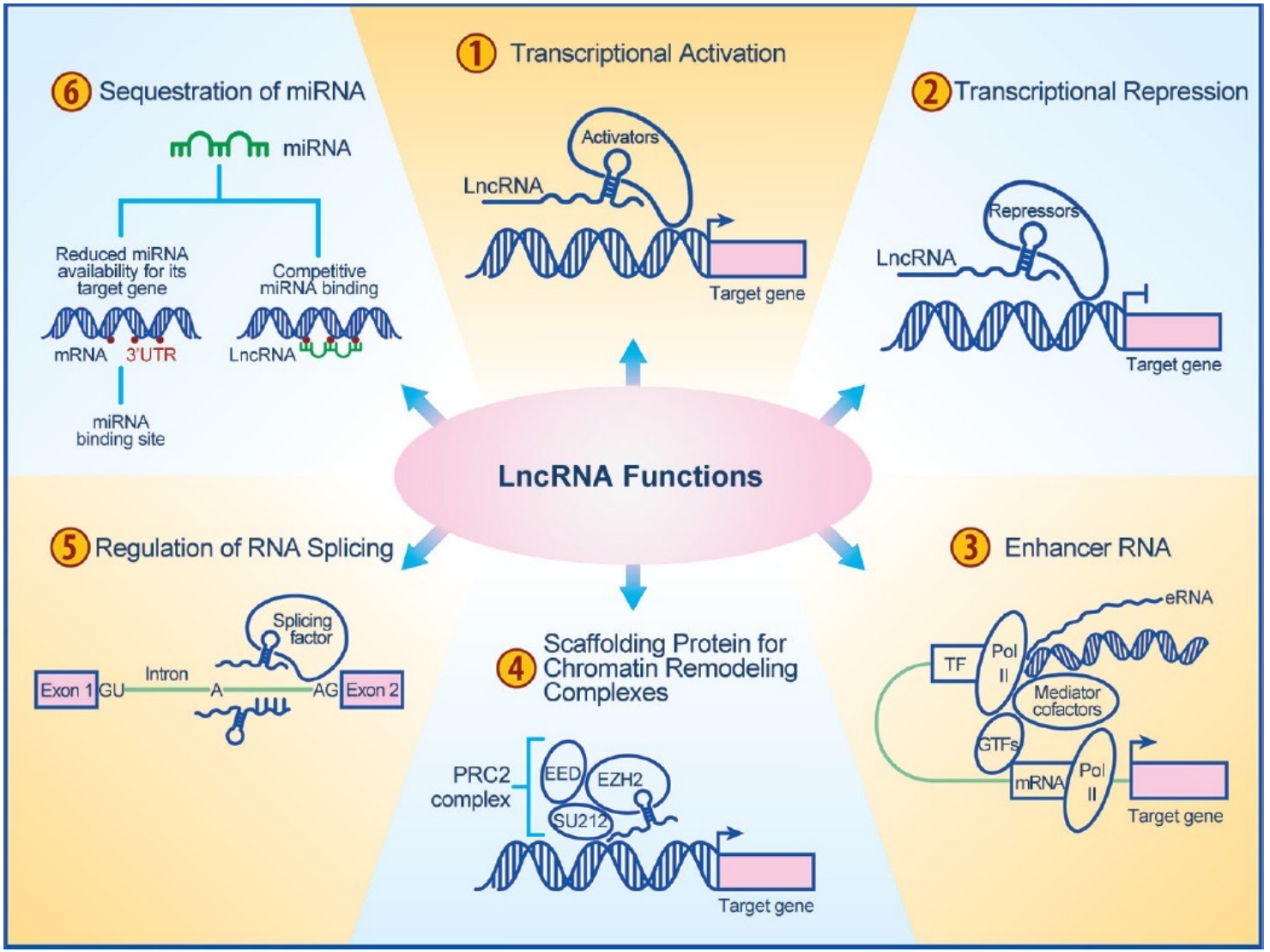
Palazzo et al, Front. Genet., 26 January 2015



RNA by mass RNA by number of molecules

Estimation of RNA levels in a mammalian cell

The different functions of lncRNAs



Malik et al, Asian J Androl. 2016

Definition of an lncRNA

LncRNA = transcript without coding potential, >200 nt, spliced, polyA+/- (Derrien et al., 2012)
Example: annotation of the human genome:

Human

Statistics about the current GENCODE Release (version 41)

The statistics derive from the gtf file that contains only the annotation of the main chromosomes.

For details about the calculation of these statistics please see the README_stats.txt file.

General stats

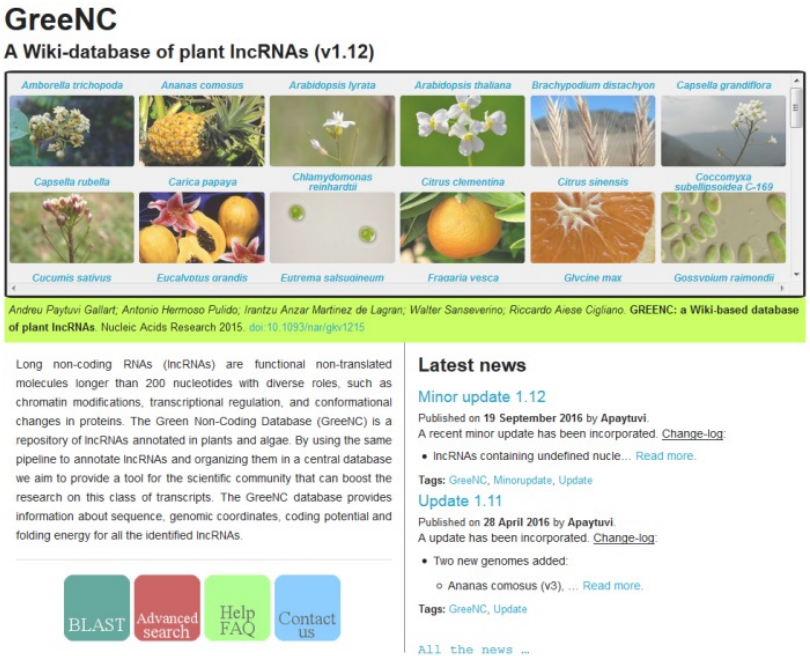
Total No of Genes	61852	Total No of Transcripts	251236
Protein-coding genes	19370	Protein-coding transcripts	88780
- readthrough genes (not included)	647	- full length protein-coding	63370
Long non-coding RNA genes	19095	- partial length protein-coding	25410
Small non-coding RNA genes	7566	Nonsense mediated decay transcripts	20933
Pseudogenes	14736	Long non-coding RNA loci transcripts	54291
- processed pseudogenes	10662		
- unprocessed pseudogenes	3573		
- unitary pseudogenes	250		
- pseudogenes	15	Total No of distinct translations	65052
Immunoglobulin/T-cell receptor gene segments		Genes that have more than one distinct translations	13614
- protein coding segments	410		
- pseudogenes	236		

https://www.gencodegenes.org/human/stats.html

Databases, resources

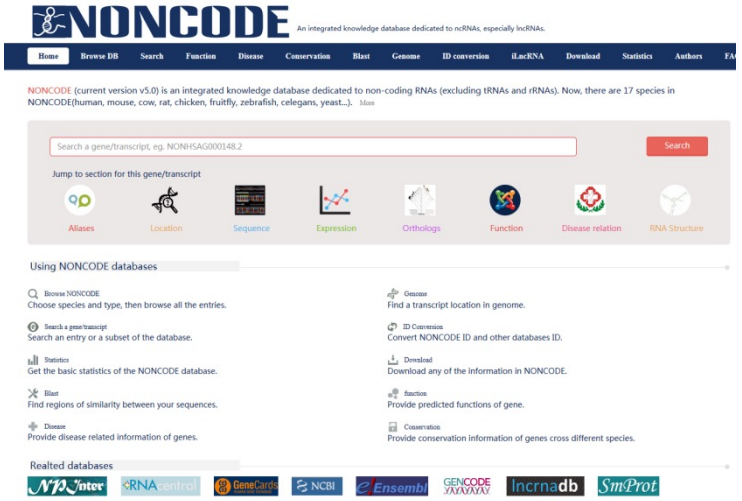
http://greenc.sequentiabiotech.com/wiki/Main_Page

A Wiki-database of plant lncRNAs



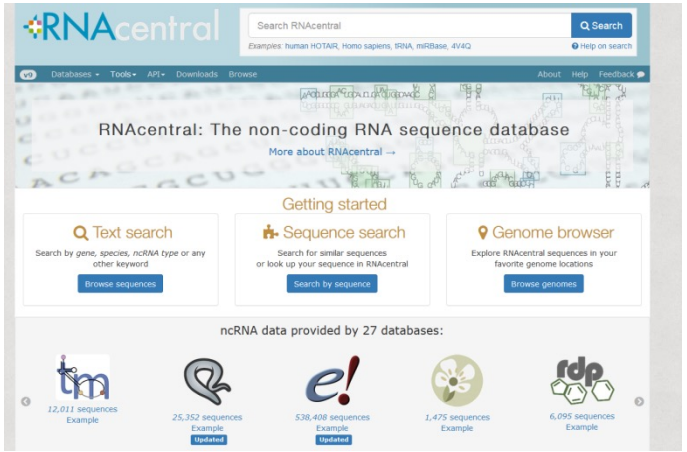
http://www.noncode.org/

database dedicated to non-coding RNAs (excluding tRNAs and rRNAs)

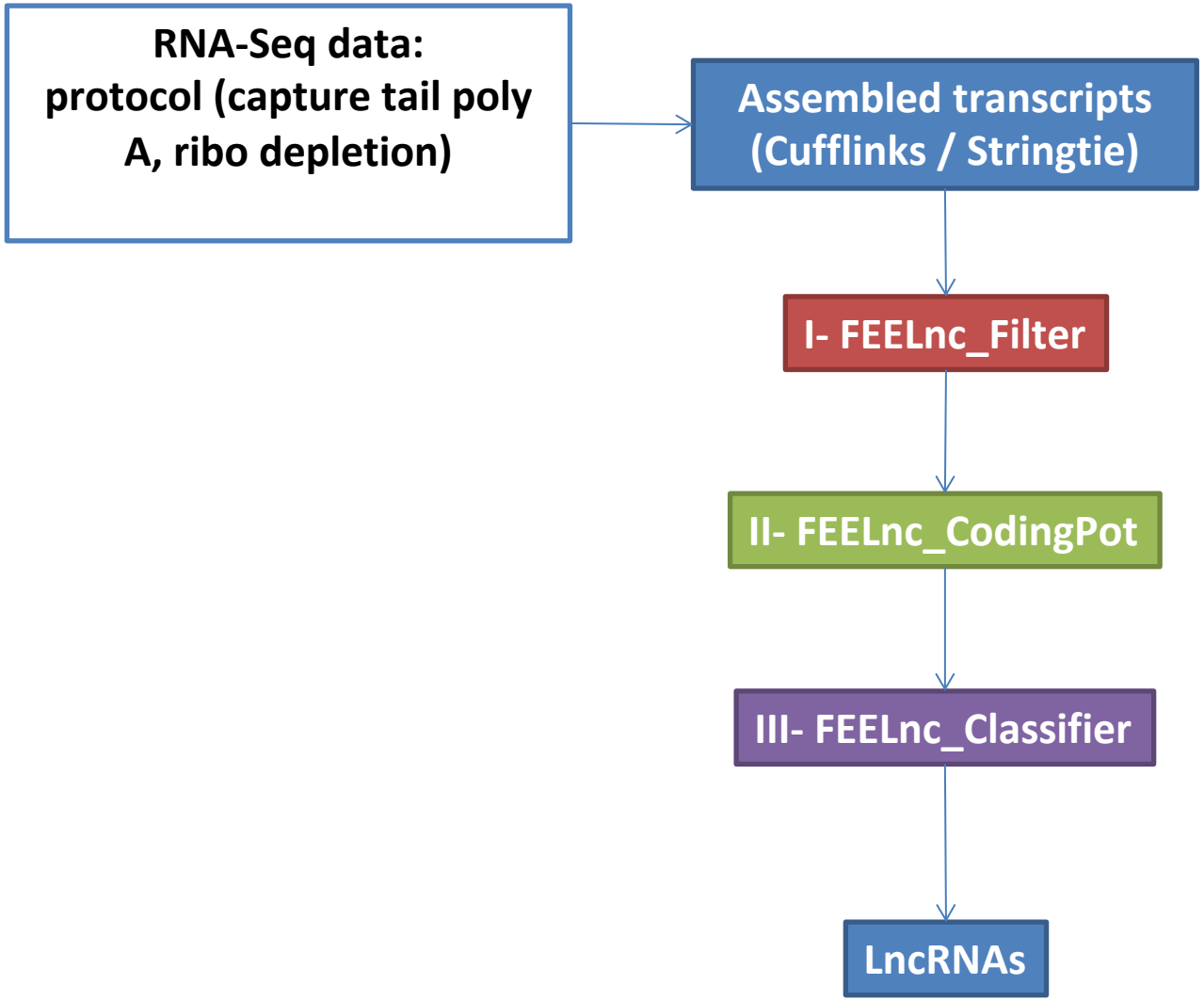


http://rnacentral.org/

The non-coding RNA sequence database



FEELnc : FlExible Extraction of LncRNAs



* FEELnc_filter.pl : Extract, filter candidate transcripts.

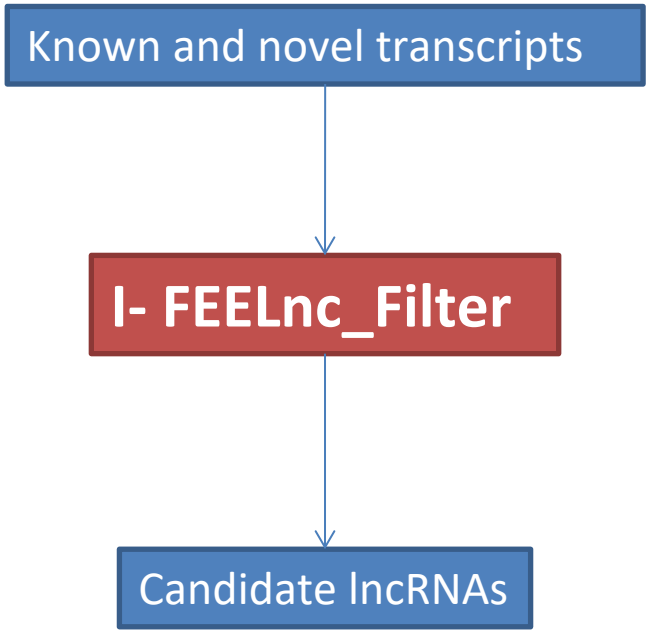
* FEELnc_codpot.pl : Compute the coding potential of candidate transcripts.

* FEELnc_classifier.pl : Classify lncRNAs based on their genomic localization wrt others transcripts.

https://github.com/tderrien/FEELnc

Wucher V, et al. FEELnc : a tool for long non-coding RNA annotation and its application to the dog transcriptome. Nucleic Acids Res. 2017 May 5 ;45(8) :e57.

I- FEELnc_Filter



Deletion of non lncRNAs:

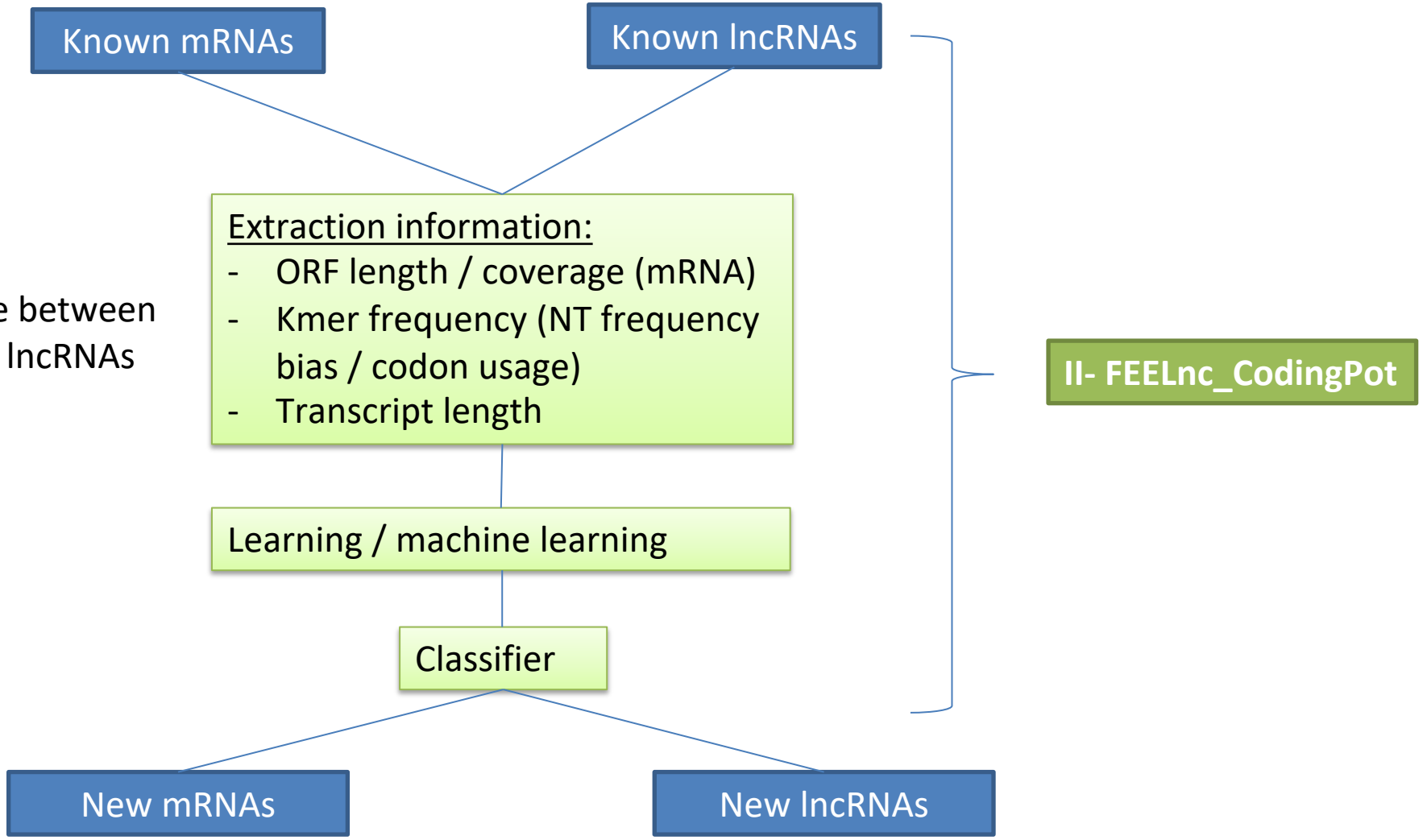
All possible types (small ncRNA, pseudogenes, ...) Severe filtering, possible loss of lncRNA.

- RNA codant
- RNA considered as potential isoforms
- < 200 bp
- Mono-exonic

II- FEELnc_CodingPot

=> Define a CPS = "Coding Potential Score" for each candidate RNA

3 criteria to discriminate between mRNAs and lncRNAs



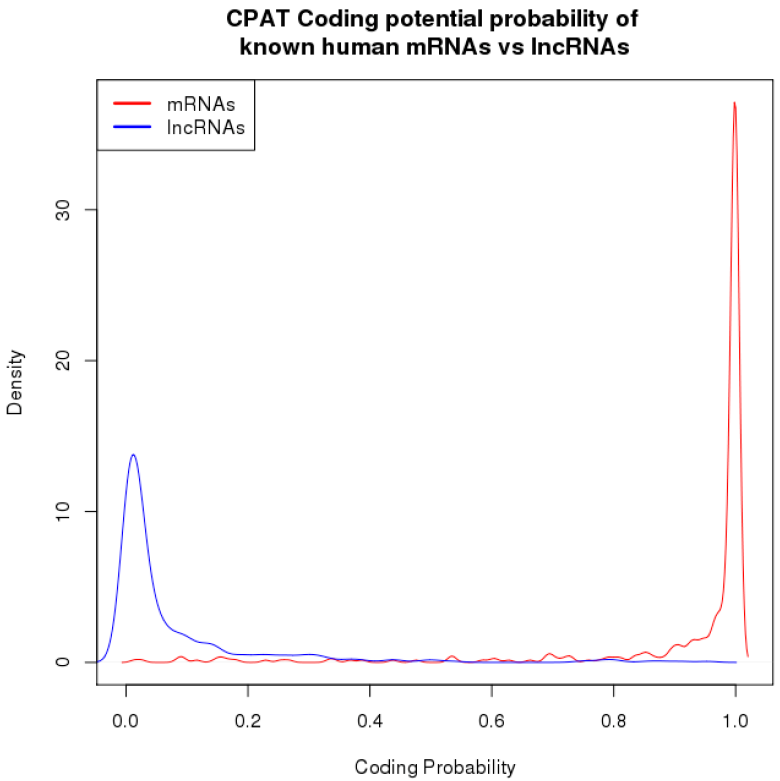
II- FEELnc_CodingPot

Extraction information:
- ORF length / coverage
- Kmer frequency
- Transcript length

RandomForest
- Easily optimized
- Adapted to unbalanced datasets
- Handles missing data

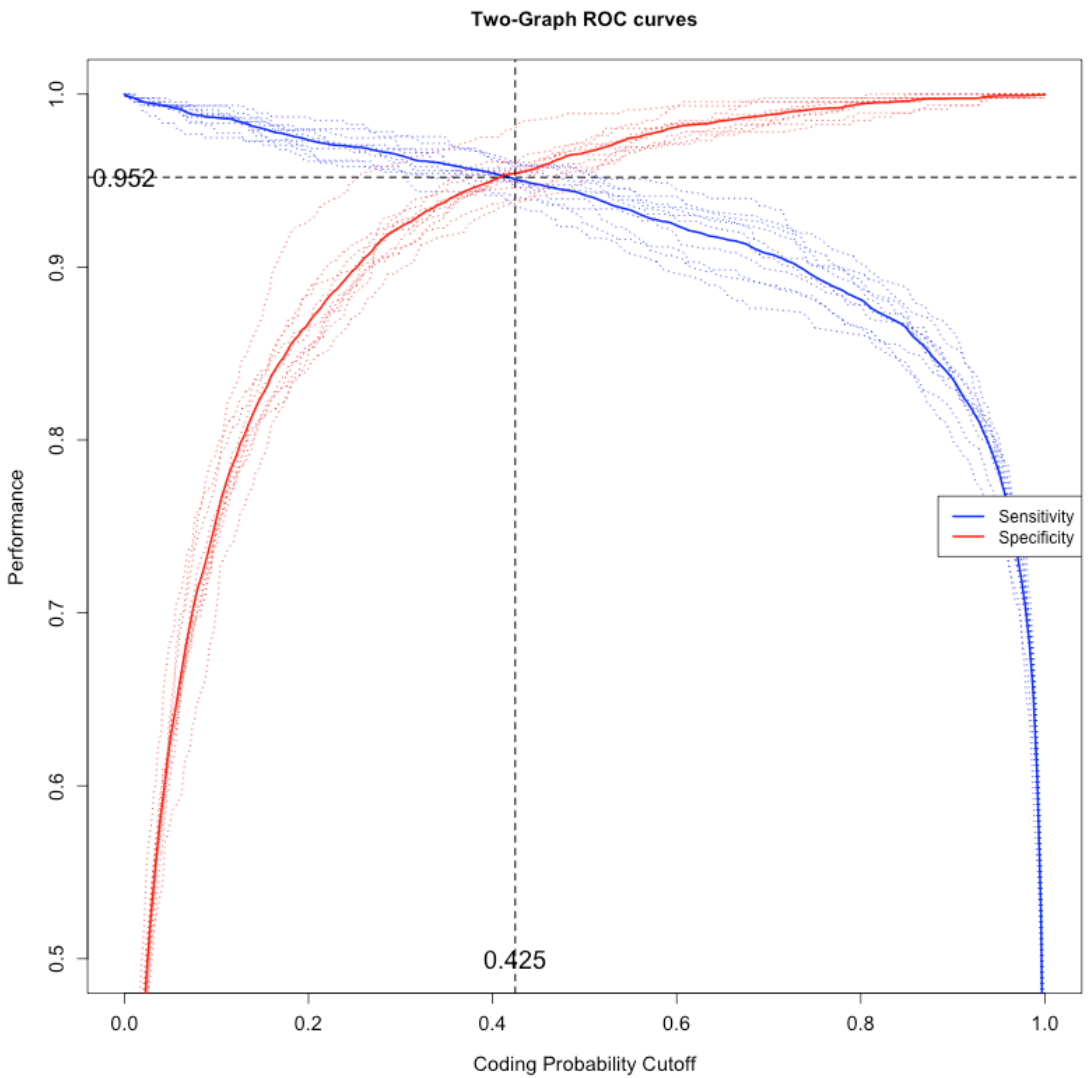
Obtaining a CPS (Coding Potential Score) for all transcripts

selection of the threshold ?



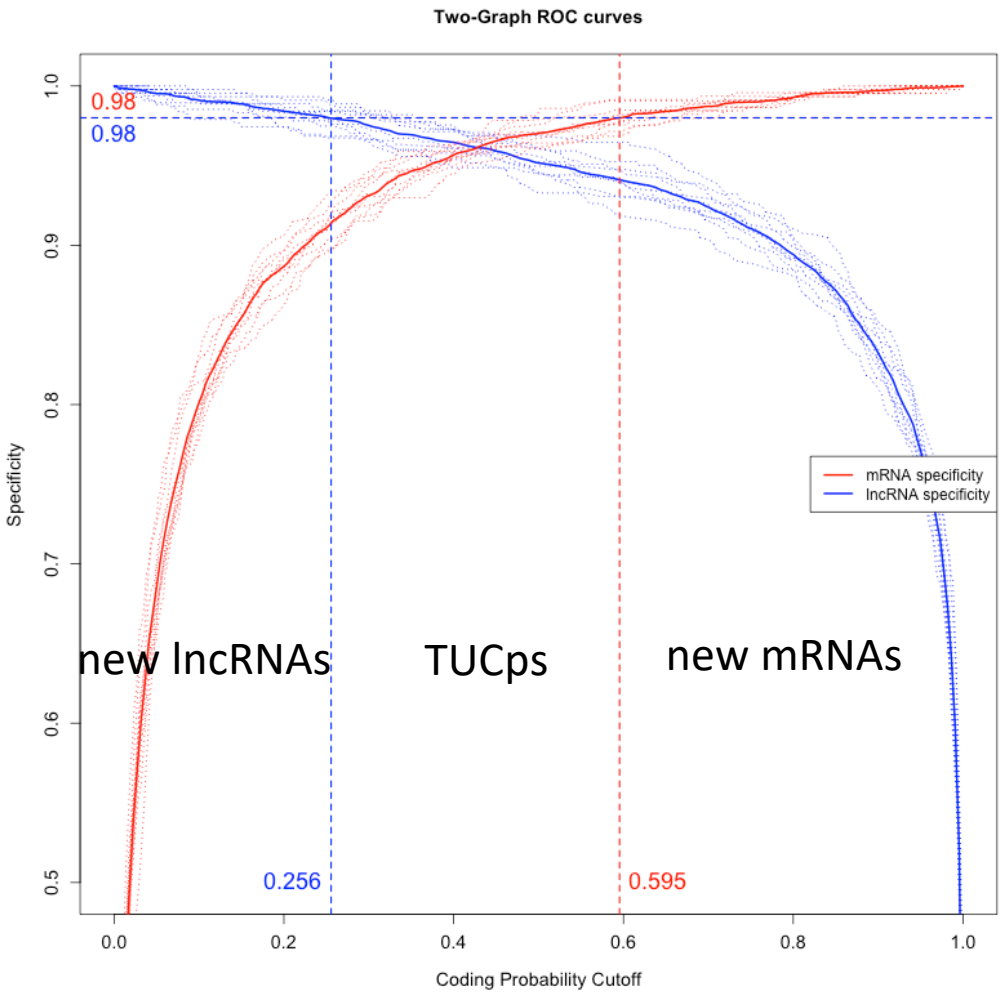
II- FEELnc_CodingPot

Choosing an optimal CPS



Cut-off = Maximization of sensitivity / specificity

II- FEELnc_CodingPot



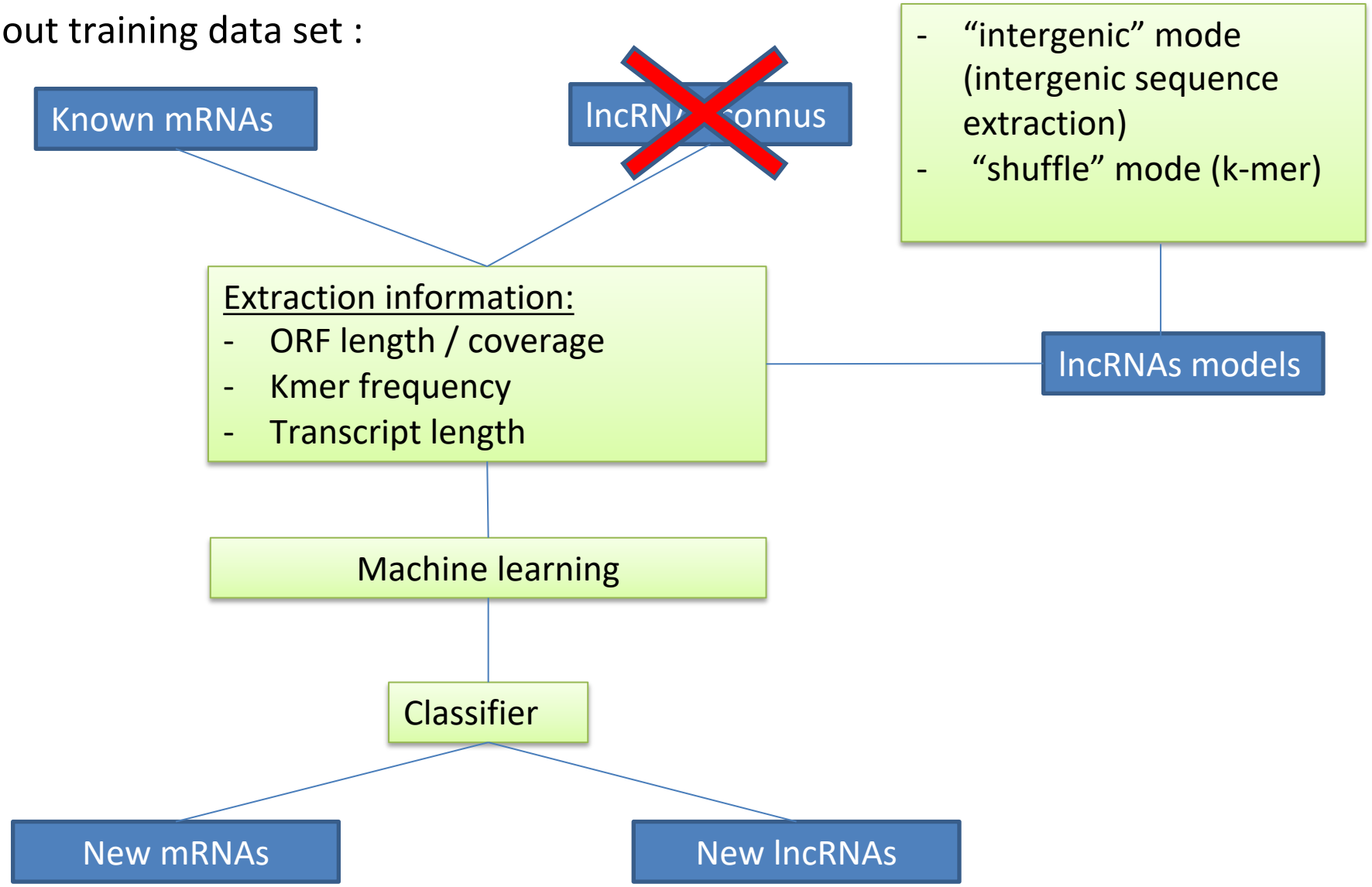
TUCPs = transcrits ambiguous

"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, Discovery and annotation of long noncoding RNAs. Nature Structural Molecular Biology, 22:5–7, 2015.

II- FEELnc_CodingPot

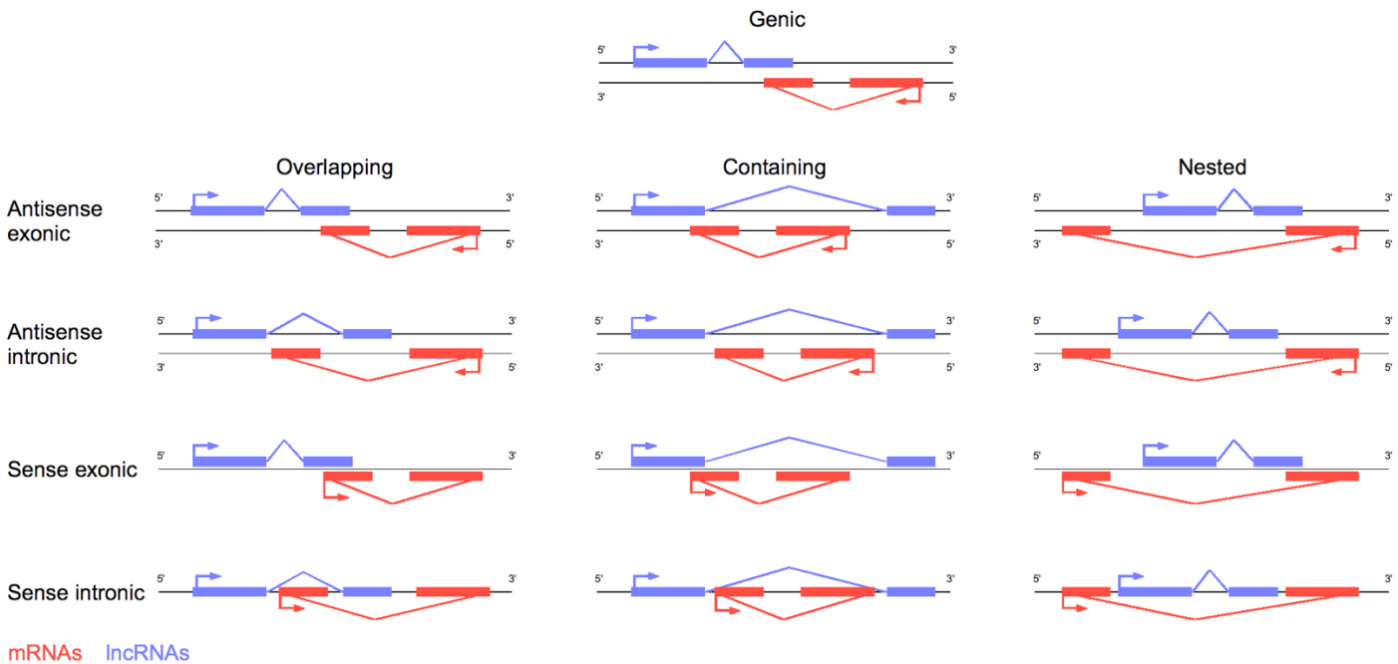
Use without training data set :



III- FEELnc_Classifier

Classification of lncRNAs according to their genomic context
Classification can help to understand **the function of lncRNA**

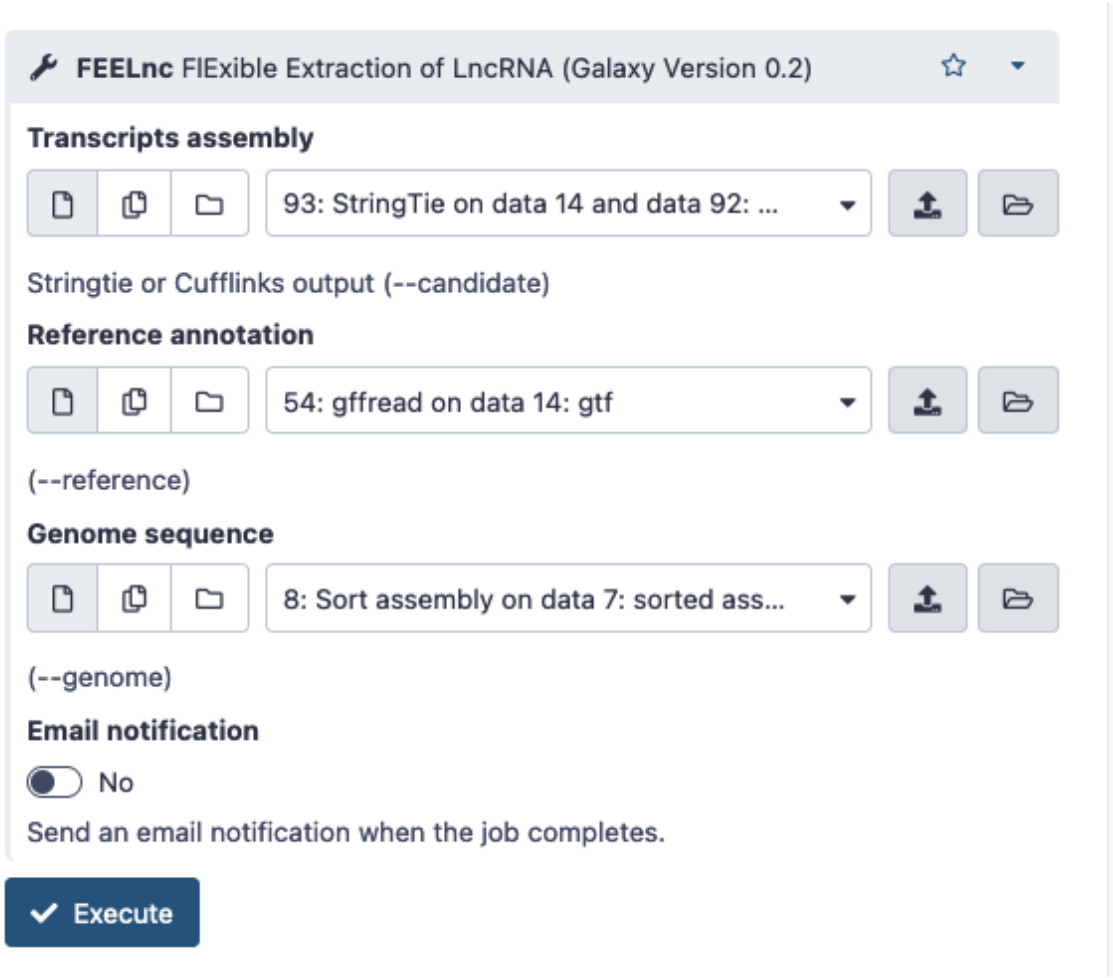
-> intergenic or **intragenic** localisation



Use in Galaxy

mRNA / lncRNA prediction : **‘FEELnc FlExible Extraction of LncRNA‘**

Inputs :



Use in Galaxy

1st Output : Tool Standard Output = summary of results

```
####################
Done: results in out_feelnc/{filter;codpot;classifier}

# Summary file:
-With_cutoff:   0.5872
-Nb_lncRNAs:    131
-Nb_mRNAs:      22
#FEELnc Classification
#lncRNA file :  lncrna : out_feelnc/codpot
//candidate_lncRNA.codpot.lncRNA.gtf
#mRNA file : /opt/galaxy-dist/database/files/002/190
/dataset_2190461.dat
#Minimal window size : 10000
#Maximal window size : 100000
#Number of lncRNA : 131
#Number of mRNA : 656
#Number of interaction : 318
#Number of lncRNA without interaction : 0
#List of lncRNA without interaction :
```

Use in Galaxy

2nds Outputs : Annotations of lncRNAs and new mRNAs (not present in the reference annotation) **in GTF format**.

Seqname	Source	Feature	Start	End	Score	Strand	Frame	Attributes
GL349630	Cufflinks	exon	2912885	2913338	1000	+	.	gene_id "CUFF.55"; transcript_id "rna1411"; FPKM "88.4493592620"; conf_hi "103.972745"; conf_lo "73.2005
GL349630	Cufflinks	exon	2913425	2913520	1000	+	.	gene_id "CUFF.55"; transcript_id "rna1411"; FPKM "88.4493592620"; conf_hi "103.972745"; conf_lo "73.2005
GL349630	Cufflinks	exon	2913770	2913820	1000	+	.	gene_id "CUFF.55"; transcript_id "rna1411"; FPKM "88.4493592620"; conf_hi "103.972745"; conf_lo "73.2005
GL349629	Cufflinks	exon	2038027	2038133	1000	+	.	gene_id "CUFF.369"; transcript_id "rna1243"; FPKM "32.3423914761"; conf_hi "36.481549"; conf_lo "28.2517
GL349629	Cufflinks	exon	2038857	2038980	1000	+	.	gene_id "CUFF.369"; transcript_id "rna1243"; FPKM "32.3423914761"; conf_hi "36.481549"; conf_lo "28.2517
GL349629	Cufflinks	exon	2039110	2039169	1000	+	.	gene_id "CUFF.369"; transcript_id "rna1243"; FPKM "32.3423914761"; conf_hi "36.481549"; conf_lo "28.2517
GL349629	Cufflinks	exon	2040268	2040772	1000	+	.	gene_id "CUFF.369"; transcript_id "rna1243"; FPKM "32.3423914761"; conf_hi "36.481549"; conf_lo "28.2517
GL349629	Cufflinks	exon	2040831	2040892	1000	+	.	gene_id "CUFF.369"; transcript_id "rna1243"; FPKM "32.3423914761"; conf_hi "36.481549"; conf_lo "28.2517
GL349629	Cufflinks	exon	2041055	2041327	1000	+	.	gene_id "CUFF.369"; transcript_id "rna1243"; FPKM "32.3423914761"; conf_hi "36.481549"; conf_lo "28.2517
GL349629	Cufflinks	exon	2042021	2042295	1000	+	.	gene_id "CUFF.369"; transcript_id "rna1243"; FPKM "32.3423914761"; conf_hi "36.481549"; conf_lo "28.2517
GL349629	Cufflinks	exon	2042385	2042660	1000	+	.	gene_id "CUFF.369"; transcript_id "rna1243"; FPKM "32.3423914761"; conf_hi "36.481549"; conf_lo "28.2517
GL349629	Cufflinks	exon	2174317	2174506	579	+	.	gene_id "CUFF.382"; transcript_id "rna1252"; FPKM "23.2928719308"; conf_hi "32.595577"; conf_lo "13.9695
GL349629	Cufflinks	exon	2174626	2174768	579	+	.	gene_id "CUFF.382"; transcript_id "rna1252"; FPKM "23.2928719308"; conf_hi "32.595577"; conf_lo "13.9695
GL349629	Cufflinks	exon	2175998	2176147	579	+	.	gene_id "CUFF.382"; transcript_id "rna1252"; FPKM "23.2928719308"; conf_hi "32.595577"; conf_lo "13.9695
GL349629	Cufflinks	exon	379490	379808	1	-	.	gene_id "CUFF.467"; transcript_id "rna1161"; FPKM "0.0000000000"; conf_hi "1.009723"; conf_lo "0.000000"
GL349629	Cufflinks	exon	379898	380064	1	-	.	gene_id "CUFF.467"; transcript_id "rna1161"; FPKM "0.0000000000"; conf_hi "1.009723"; conf_lo "0.000000"
GL349629	Cufflinks	exon	399685	399769	1	-	.	gene_id "CUFF.467"; transcript_id "rna1161"; FPKM "0.0000000000"; conf_hi "1.009723"; conf_lo "0.000000"
GL349686	Cufflinks	exon	294130	294597	770	-	.	gene_id "CUFF.401"; transcript_id "CUFF.401.2"; FPKM "2.2583944874"; conf_hi "3.787267"; conf_lo "0.78901
GL349686	Cufflinks	exon	294655	294851	770	-	.	gene_id "CUFF.401"; transcript_id "CUFF.401.2"; FPKM "2.2583944874"; conf_hi "3.787267"; conf_lo "0.78901
GL349686	Cufflinks	exon	295825	295921	770	-	.	gene_id "CUFF.401"; transcript_id "CUFF.401.2"; FPKM "2.2583944874"; conf_hi "3.787267"; conf_lo "0.78901
GL349686	Cufflinks	exon	296627	296851	770	-	.	gene_id "CUFF.401"; transcript_id "CUFF.401.2"; FPKM "2.2583944874"; conf_hi "3.787267"; conf_lo "0.78901
GL349686	Cufflinks	exon	296922	297067	770	-	.	gene_id "CUFF.401"; transcript_id "CUFF.401.2"; FPKM "2.2583944874"; conf_hi "3.787267"; conf_lo "0.78901
GL349686	Cufflinks	exon	298640	298863	770	-	.	gene_id "CUFF.401"; transcript_id "CUFF.401.2"; FPKM "2.2583944874"; conf_hi "3.787267"; conf_lo "0.78901
GL350496	Cufflinks	exon	15	1922	1000	+	.	gene_id "CUFF.201"; transcript_id "CUFF.201.1"; FPKM "78.1897922262"; conf_hi "83.663927"; conf_lo "72.72
GL350496	Cufflinks	exon	1985	2200	1000	+	.	gene_id "CUFF.201"; transcript_id "CUFF.201.1"; FPKM "78.1897922262"; conf_hi "83.663927"; conf_lo "72.72

Use in Galaxy

3rd Output : tabulated file from « Classifier »

isBest	lncRNA_gene	lncRNA_transcript	partnerRNA_gene	partnerRNA_transcript	direction	type	distance	subtype	location
1	CUFF.499	CUFF.499.1	ACYPI24120	NM_001293496.1	strand_unknow	genic	0	containing	exonic
0	CUFF.499	CUFF.499.1	LOC100167693	XM_001952323.4	strand_unknow	intergenic	7034	unknow strand(s)	upstream
0	CUFF.499	CUFF.499.1	Rps12	NM_001126187.2	strand_unknow	intergenic	6004	unknow strand(s)	downstream
1	CUFF.406	CUFF.406.1	LOC100161406	XM_008180842.2	strand_unknow	intergenic	879	unknow strand(s)	upstream
0	CUFF.406	CUFF.406.1	LOC107882602	XM_016801190.1	strand_unknow	intergenic	4814	unknow strand(s)	downstream
1	CUFF.23	CUFF.23.2	LOC103309142	XM_008183847.2	strand_unknow	intergenic	12094	unknow strand(s)	downstream
0	CUFF.23	CUFF.23.2	LOC103309144	XM_008183856.2	strand_unknow	intergenic	17126	unknow strand(s)	downstream
1	CUFF.465	CUFF.465.1	LOC100569545	XM_003245770.3	strand_unknow	genic	0	overlapping	exonic
0	CUFF.465	CUFF.465.1	LOC100164094	XM_001944602.4	strand_unknow	genic	0	nested	intronic
0	CUFF.465	CUFF.465.1	LOC100164094	XM_016805959.1	strand_unknow	genic	0	nested	intronic
0	CUFF.465	CUFF.465.1	LOC100163248	XM_001948643.4	strand_unknow	intergenic	200	unknow strand(s)	upstream
1	CUFF.411	CUFF.411.1	LOC103308104	XM_008180846.1	strand_unknow	intergenic	1469	unknow strand(s)	downstream
0	CUFF.411	CUFF.411.1	LOC100572330	XM_016801195.1	strand_unknow	intergenic	4212	unknow strand(s)	upstream
1	CUFF.68	rna1419	LOC107883528	XM_016803708.1	sense	intergenic	2428	same_strand	downstream
1	CUFF.228	CUFF.228.1	LOC100167418	XM_016807463.1	sense	intergenic	66	same_strand	downstream
0	CUFF.228	CUFF.228.1	LOC100168306	XM_016807621.1	sense	intergenic	9003	same_strand	upstream
0	CUFF.228	CUFF.228.1	LOC100571068	XM_016807628.1	sense	intergenic	8130	same_strand	downstream
0	CUFF.228	CUFF.228.1	LOC100168306	XM_016807619.1	sense	intergenic	9006	same_strand	upstream
0	CUFF.228	CUFF.228.1	LOC100168306	XM_001943414.4	sense	intergenic	9006	same_strand	upstream
1	CUFF.31	rna1352	LOC100159340	XM_001945434.4	sense	intergenic	6613	same_strand	upstream
1	CUFF.249	rna2476	LOC107884832	XR_001680174.1	antisense	genic	0	containing	intronic
0	CUFF.249	rna2476	LOC100158707	XM_001947526.4	antisense	intergenic	909	divergent	upstream
0	CUFF.249	rna2476	LOC107884832	XM_016807755.1	antisense	genic	0	containing	intronic
1	CUFF.145	CUFF.145.1	LOC103307681	XM_016802187.1	strand_unknow	intergenic	975	unknow strand(s)	downstream
0	CUFF.145	CUFF.145.1	LOC103307681	XM_008182050.2	strand_unknow	intergenic	975	unknow strand(s)	downstream
0	CUFF.145	CUFF.145.1	LOC103307681	XM_016802186.1	strand_unknow	intergenic	975	unknow strand(s)	downstream
0	CUFF.145	CUFF.145.1	LOC100159380	XM_001951160.4	strand_unknow	intergenic	4768	unknow strand(s)	upstream
1	CUFF.434	rna1294	LOC100168462	XM_001943125.4	sense	intergenic	1808	same_strand	upstream
0	CUFF.434	rna1294	LOC100159563	XM_001943055.4	sense	intergenic	9450	same_strand	upstream
1	CUFF.513	CUFF.513.1	LOC100165064	XM_001952350.4	antisense	genic	0	containing	exonic
0	CUFF.513	CUFF.513.1	LOC100162018	XM_008184018.2	antisense	intergenic	8991	convergent	downstream
0	CUFF.513	CUFF.513.1	LOC100162018	XM_016803932.1	antisense	intergenic	8991	convergent	downstream
0	CUFF.513	CUFF.513.1	LOC100162018	XM_001947939.4	antisense	intergenic	8991	convergent	downstream
0	CUFF.513	CUFF.513.1	LOC100162018	XM_008184013.2	antisense	intergenic	8991	convergent	downstream
0	CUFF.513	CUFF.513.1	LOC100162018	XM_008184006.2	antisense	intergenic	8991	convergent	downstream
1	CUFF.501	rna1393	LOC100568818	XM_003240355.3	sense	intergenic	20807	same_strand	downstream
0	CUFF.501	rna1393	ACYPI24120	NM_001293496.1	sense	intergenic	22303	same_strand	upstream
1	CUFF.370	rna1237	LOC103309036	XM_016803131.1	sense	intergenic	288	same_strand	downstream
1	CUFF.118	rna17294	LOC100160162	XM_008185501.2	antisense	intergenic	843	convergent	downstream
0	CUFF.118	rna17294	LOC100160162	XM_001950678.4	antisense	intergenic	843	convergent	downstream
0	CUFF.118	rna17294	LOC100160162	XM_003244927.3	antisense	intergenic	843	convergent	downstream
0	CUFF.118	rna17294	LOC103309617	XM_008185502.2	antisense	intergenic	3133	divergent	upstream
0	CUFF.118	rna17294	LOC103309617	XM_016804903.1	antisense	intergenic	3133	divergent	upstream
1	CUFF.231	CUFF.231.1	LOC100160756	XM_001943193.4	sense	intergenic	19741	same_strand	upstream
1	CUFF.391	CUFF.391.1	LOC103308096	XM_008180837.1	sense	intergenic	3309	same_strand	downstream
1	CUFF.382	CUFF.382.1	LOC100166241	XM_001950566.4	sense	intergenic	9194	same_strand	upstream
1	CUFF.454	CUFF.454.1	LOC100159416	XM_016803814.1	strand_unknow	intergenic	1976	unknow strand(s)	upstream
0	CUFF.454	CUFF.454.1	Su(var)3-9	NM_001126162.2	strand_unknow	intergenic	9120	unknow strand(s)	downstream
0	CUFF.454	CUFF.454.1	LOC100573350	XM_008184115.2	strand_unknow	intergenic	6746	unknow strand(s)	downstream
1	CUFF.401	rna6096	LOC100571289	XM_003241763.3	sense	intergenic	83405	same_strand	downstream