Long non-coding RNAs (lncRNAs) annotation

EBAII Assembly & Annotation - Roscoff Juin 2024

Romane LIBOUBAN
GenOuest platform, Rennes

Importance et fonction des ARN non codants

- 80 % des variants associés aux maladies ne sont pas situées sur les gènes codant pour les protéines. (Manolio et al, Hindorrf et al)
- Plus de 60 % du génome humain est transcrit en ARN, mais seuls 2 % seront traduits en protéines.

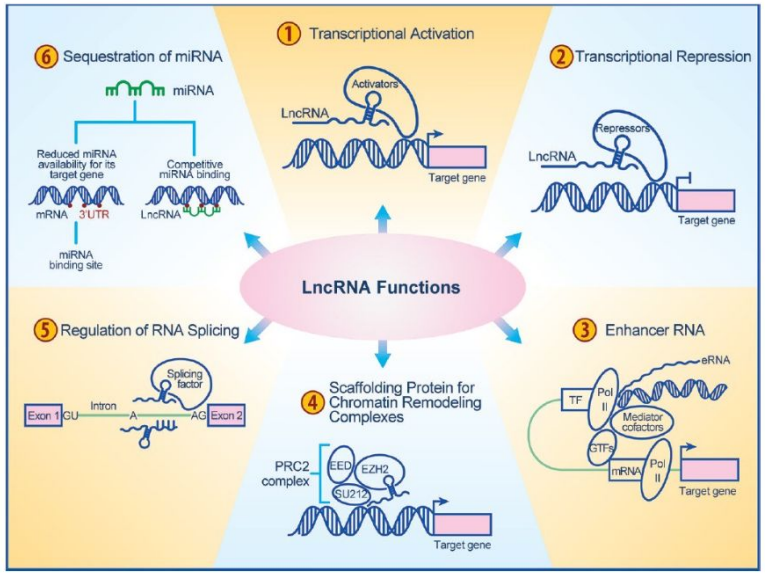
→ **Nécessité d'annoter les ARN non codants pour améliorer la compréhension du génotype et du phénotype**

Types d'ARN et leur rôle

- Synthèse des protéines : ARNm, ARNt, ARNr
- Régulation de l'expression des gènes : miARN, siARN, lncARN
- Modification de l'ARN : snoRNA, gRNA
- Protection du génome : piRNA

Qu'est-ce qu'un long ARN non codant ?

- Transcript ne codant pas pour une protéine
- Divers rôles dans
  - Régulation des gènes
  - Modifications d'autres ARNs
  - Autres processus cellulaires essentiels
- Taille : > 200 nucléotides
- Polyadénylation



Malik et al, Asian J Androl. 2016

Annotation du génome humain

Human

Statistics about the current GENCODE Release (version 46)

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	63086	Total No of Transcripts	254070
Protein-coding genes	19411	Protein-coding transcripts	89581
- readthrough genes (not included)	654	- full length protein-coding	64695
Long non-coding RNA genes	20310	- partial length protein-coding	24886
Small non-coding RNA genes	7565	Nonsense mediated decay transcripts	21774
Pseudogenes	14716	Long non-coding RNA loci transcripts	59927
- processed pseudogenes	10657		
- unprocessed pseudogenes	3564		
- unitary pseudogenes	258		
Immunoglobulin/T-cell receptor gene segments		Total No of distinct translations	65650
- protein coding segments	411	Genes that have more than one distinct translations	13620
- pseudogenes	237		

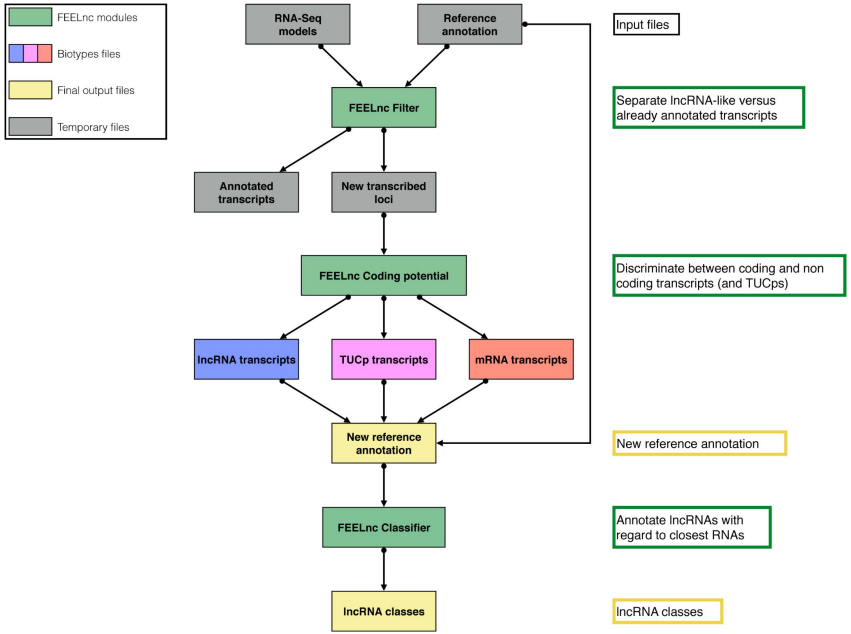
https://www.gencodegenes.org/human/stats.html

Databases et ressources

- noncode (http://www.noncode.org/)
  - ARNnc, à l'exclusion des ARNt et ARNr
- GreyNC (http://greenc.sequentiabiotech.com/wiki/Main_Page)
  - ARNnc des plantes
- rnacentral (http://rnacentral.org/)
  - ARN non codants

FEELnc: **FlE**xible **E**xtraction of **Lnc**RNAs

- Pipeline d'annotation et de classification des longs ARN non codants
- Sur la base des transcrits reconstruits à partir des données RNA-seq
- 3 modules = 3 étapes
  - Filter
  - codpot
  - classifier



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

Module filter

- Extraction et filtrage des transcripts indésirables et les transcrits chevauchant les exons de l'annotation de référence
  - Particulièrement les exons codant pour les protéines → probablement des isoformes
  - Possibilité de paramétrer le % de chevauchement et le biotype de transcription
- Filtrage sévère
  - < 200 nucléotides (par défaut)
  - Transcripts mono-exoniques
  - Perte potentielle d'ARN non codants
  - Maintien des ARNs > 200 nucléotides
  - **Suppression des ARN non lnc**

Module codpot (coding potential)

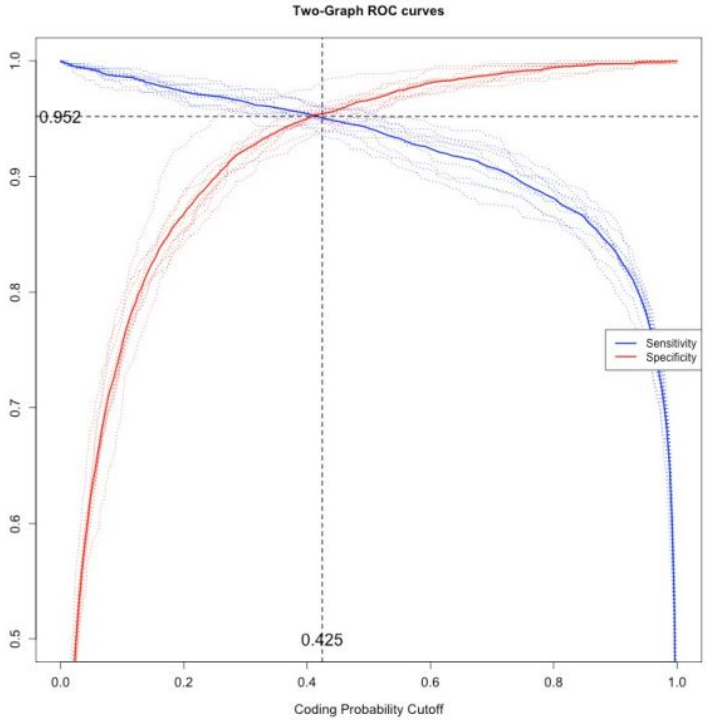
- Calculer le potentiel de codage des transcripts (CPS : Coding Potential Score)
  - 0 (ARN nc) ou 1 (ARNm)
- Distinguer davantage les lncRNA des ARN codants potentiels
- Input
  - Transcrits présélectionnées et filtrés (issus du module Filter)
- Output
  - Classification des transcrits en tant que lncRNA ou ARNm
  - Basée sur le CPS

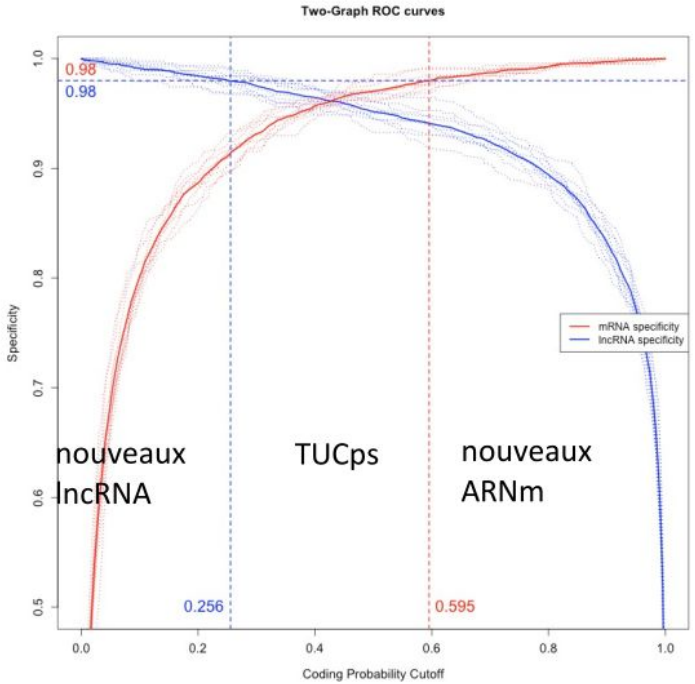
Module codpot (coding potential)

- Exploite les propriétés intrinsèques des séquences d’entrée
  - Couverture ORF
  - Taille des ARNm
  - Fréquences multi-k-mer (KIS) entre les ARNm et les ARNlnc
- Random Forest
  - Facilement optimisé
  - Adapté aux ensembles de données déséquilibrées
  - Traitement des données manquantes
  - Permet de “prédire” si une séquence est codante ou non
  - Package ROCR + Validation croisée 10 fois

Module codpot (coding potential)

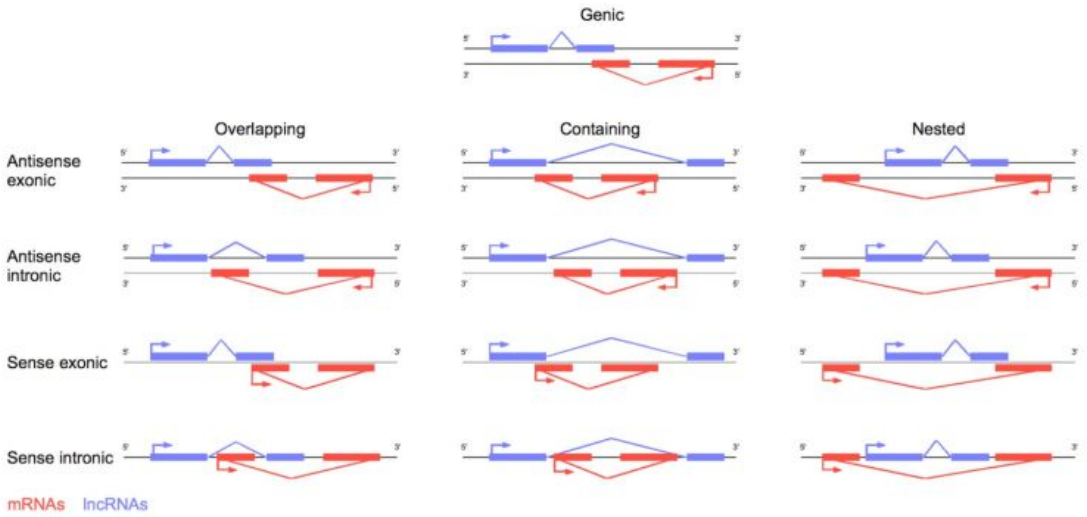
Choix d'un CPS optimale





Module classifier

- Classer les nouveaux lncRNA en fonction de la localisation et de la direction de la transcription des transcrits d'ARN proximaux
- Localisation intergénique ou intragénique



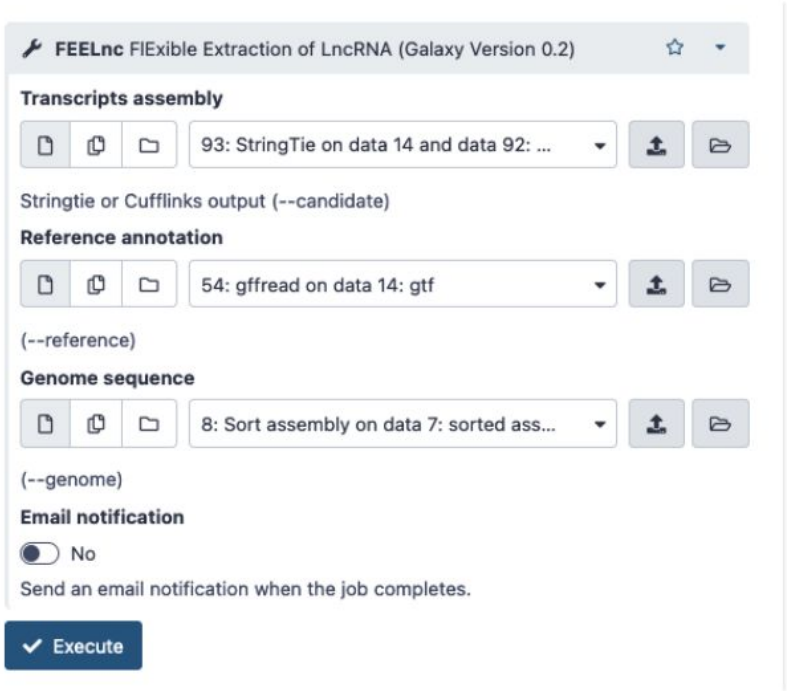
Références

- Github : https://github.com/tderrien/FEELnc?tab=readme-ov-file#input-files
- Article : https://academic.oup.com/nar/article/45/8/e57/2798184

Using FEELnc on Galaxy

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

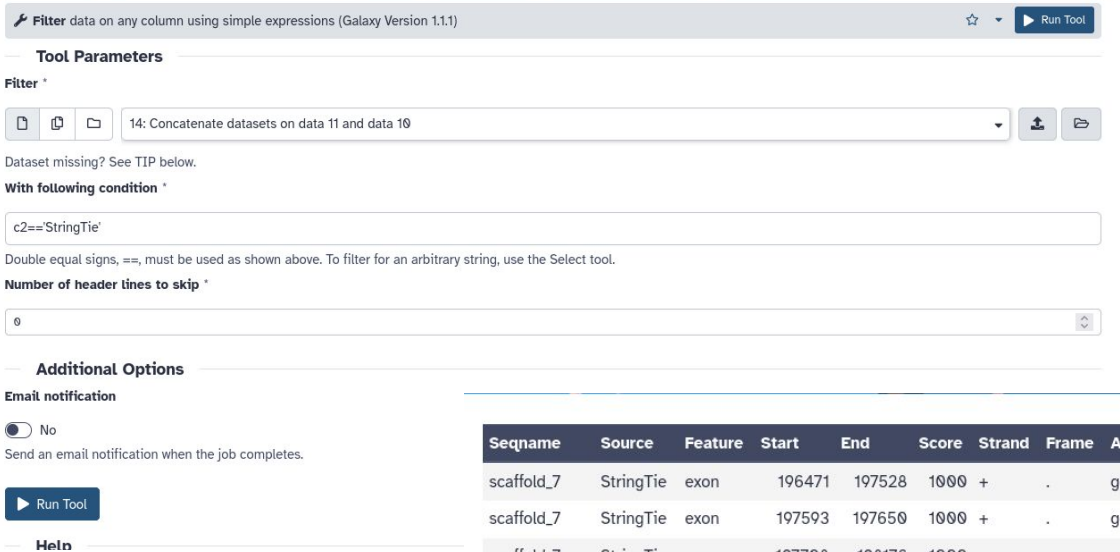
Inputs :



TP

https://training.galaxyproject.org/topics/genome-annotation/tutorials/lncrna/tutorial.html

Petits tips - Visualisation des lncRNA



Seqname	Source	Feature	Start	End	Score	Strand	Frame	Attributes
scaffold_7	StringTie	exon	196471	197528	1000	+	.	gene_id "STRG.1364"; transcript_id "STRG.1364.2"; cov "10.865439"; exon_number "1";
scaffold_7	StringTie	exon	197593	197650	1000	+	.	gene_id "STRG.1364"; transcript_id "STRG.1364.2"; cov "14.245408"; exon_number "2";
scaffold_7	StringTie	exon	197708	198176	1000	+	.	gene_id "STRG.1364"; transcript_id "STRG.1364.2"; cov "5.865661"; exon_number "3";
scaffold_79	StringTie	exon	166631	166978	1000	-	.	gene_id "STRG.8216"; transcript_id "STRG.8216.1"; cov "425.729889"; exon_number "1";
scaffold_79	StringTie	exon	167047	168177	1000	-	.	gene_id "STRG.8216"; transcript_id "STRG.8216.1"; cov "183.999115"; exon_number "2";
scaffold_94	StringTie	exon	106397	106745	1000	-	.	gene_id "STRG.8946"; transcript_id "STRG.8946.1"; cov "5279.410645"; exon_number "1";
scaffold_94	StringTie	exon	106813	107237	1000	-	.	gene_id "STRG.8946"; transcript_id "STRG.8946.1"; cov "11746.810547"; exon_number "2";
scaffold_126	StringTie	exon	46642	47035	1000	+	.	gene_id "STRG.9988"; transcript_id "STRG.9988.1"; cov "221.967010"; exon_number "1";
scaffold_126	StringTie	exon	47096	47163	1000	+	.	gene_id "STRG.9988"; transcript_id "STRG.9988.1"; cov "426.220581"; exon_number "2";