Annotation of lncRNAs – FEELnc: FlExible Extraction of LncRNAs

Thomas DERRIEN, Valentin Wucher
(tderrien@univ-rennes1.fr)

Canine Genetics Group
IGDR : Institute of Genetics and Development of Rennes
CNRS - UMR6290 - Université de Rennes 1

Fabrice Legeai
(fabrice.legeai@inra.fr)

BioInformatics Platform for Agroecosystems Arthropods
IGEPP : Institute for Genetics, Environment and Plant Protection
INRA

Non-coding RNAs

- **80% of the variants** associated with disease (by GWAS) are localized **outside of protein-coding genes** (Manolio et al., Hindorrf et al.)
- **>60% of the human genome** is transcribed into RNAs (~75% by primary transcripts) **with only 2% corresponding to proteins...** (Human ENCODE Consortium; Djebali et al. 2012, Mouse ENCODE Consortium; 2015)

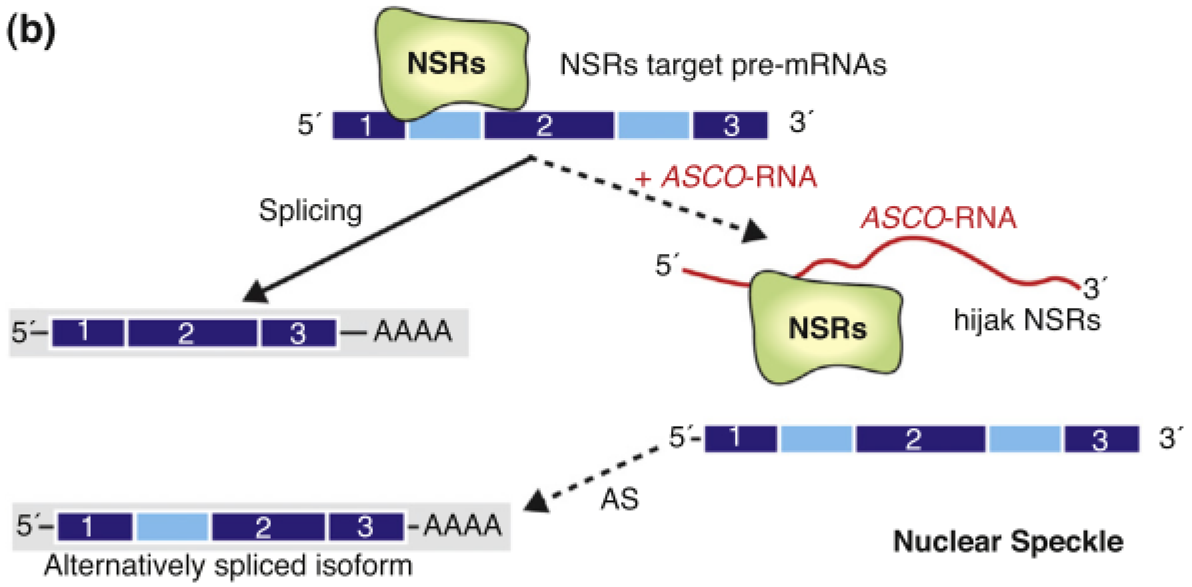
=> Need to annotate ncRNAs to ease the interpretation of genotype to phenotype relationships

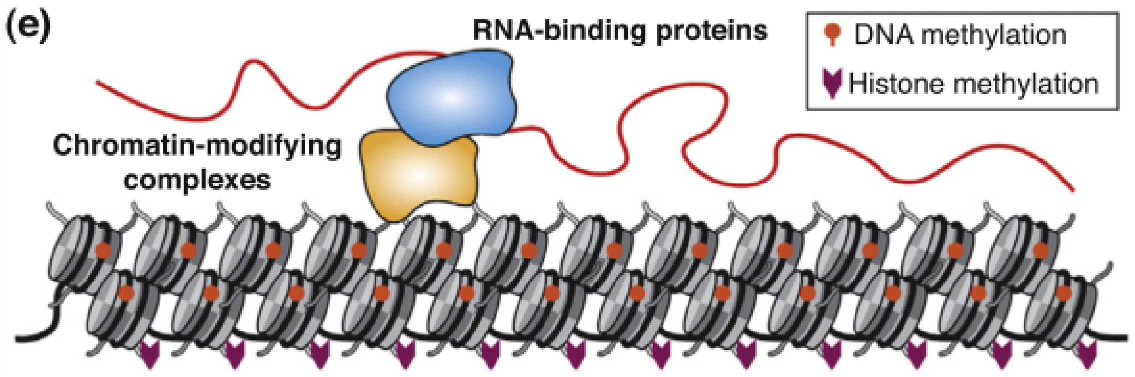
Non-coding RNAs

- **80% of the variants** associated with disease (by GWAS) are localized **outside of protein-coding genes** (Manolio et al., Hindorrf et al.)
- **>60% of the human genome** is transcribed into RNAs (~75% by primary transcripts) **with only 2% corresponding to proteins...** (Human ENCODE Consortium; Djebali et al. 2012, Mouse ENCODE Consortium; 2015)

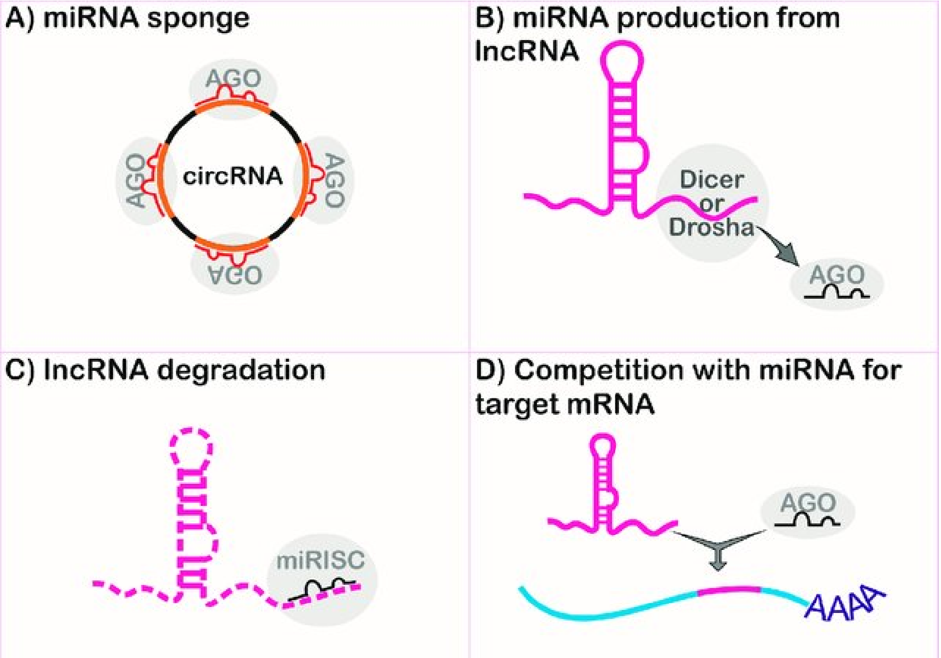
(bio)typeW	functions
mRNAs	many..
miRNAs	Regulation of gene expression
siRNAs	RNA interference pathway
snoRNAs	Chemical modification of rRNA, tRNAs and small RNAs
piRNAs	transposon defense - regulate euchromatin formation
snRNA	splicing, regulation of TFs, telomere stability...
long ncRNAs **(Xist, H19, Hotair..)**	**regulation of mRNAs expression, X chromosome inactivation, imprinting.**

LncRNAs Functions in plant epigenetics

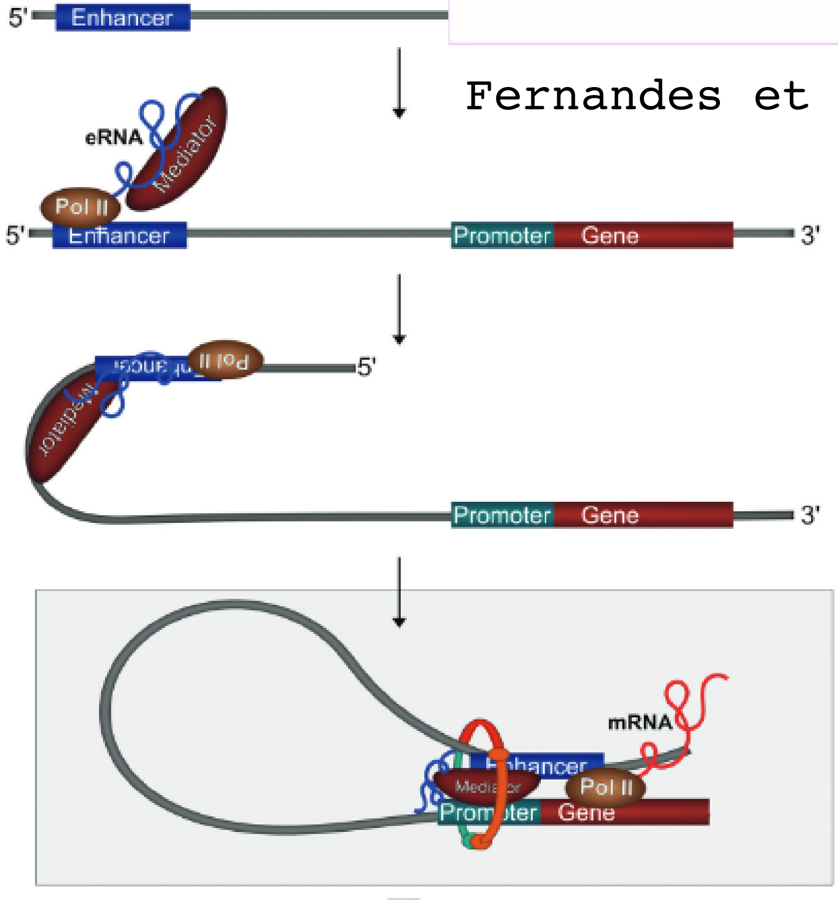




Chekanova 2015
Current Opinion in Plant Biology.

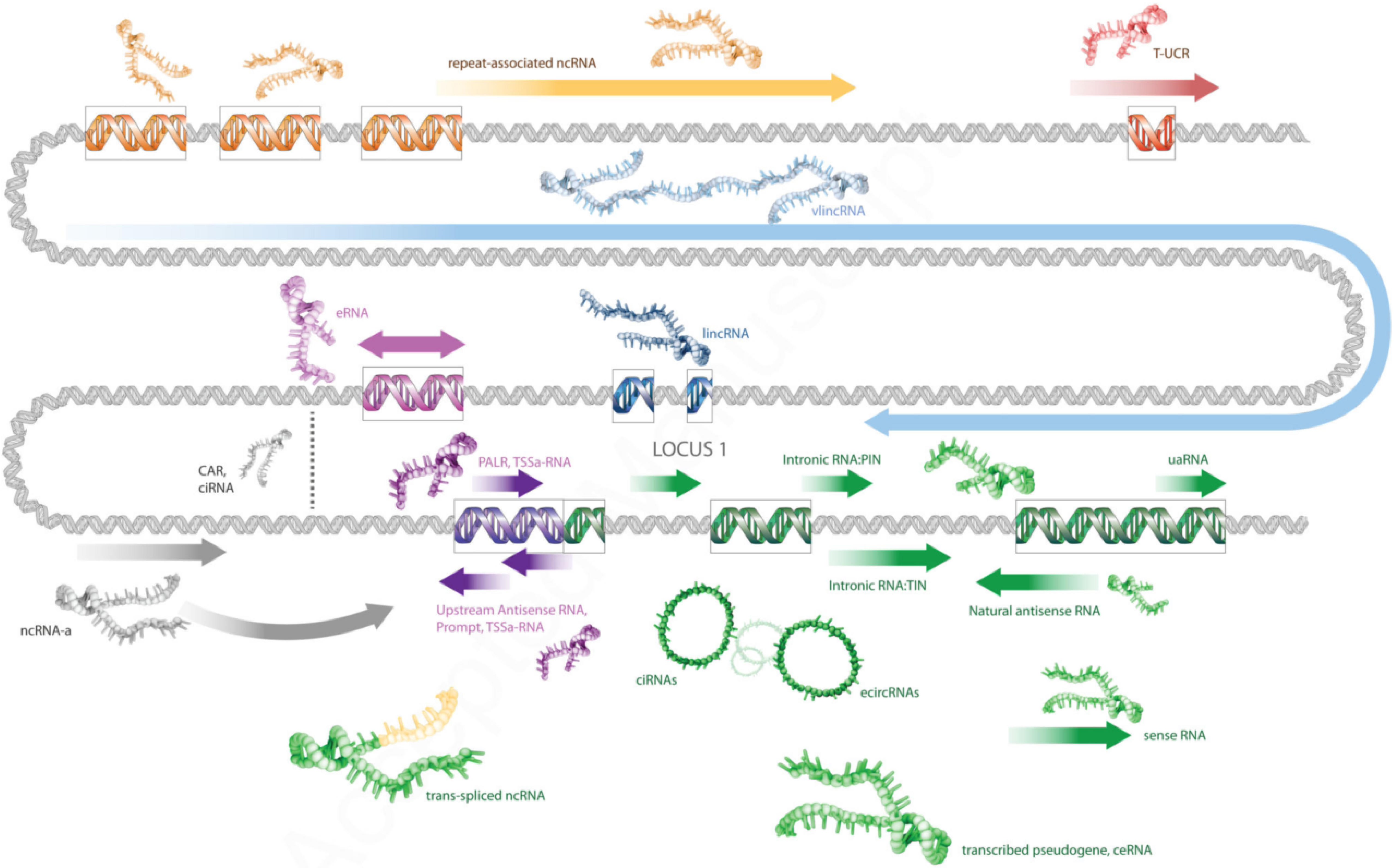


Fernandes et al. 2019 ncRNA



Shibayama et al.2014 Transcription.

The several classes of lncRNAs



St Laurent et al. Trends in Genetics 2015

lncRNAs versus mRNAs

lncRNAs are not mRNAs…

- No coding potential => small ORFs
- Less conserved (20% bw Human-Mouse)
- Lowly expressed and more tissue-specific
- Overlap many Transposable Elements (TEs)

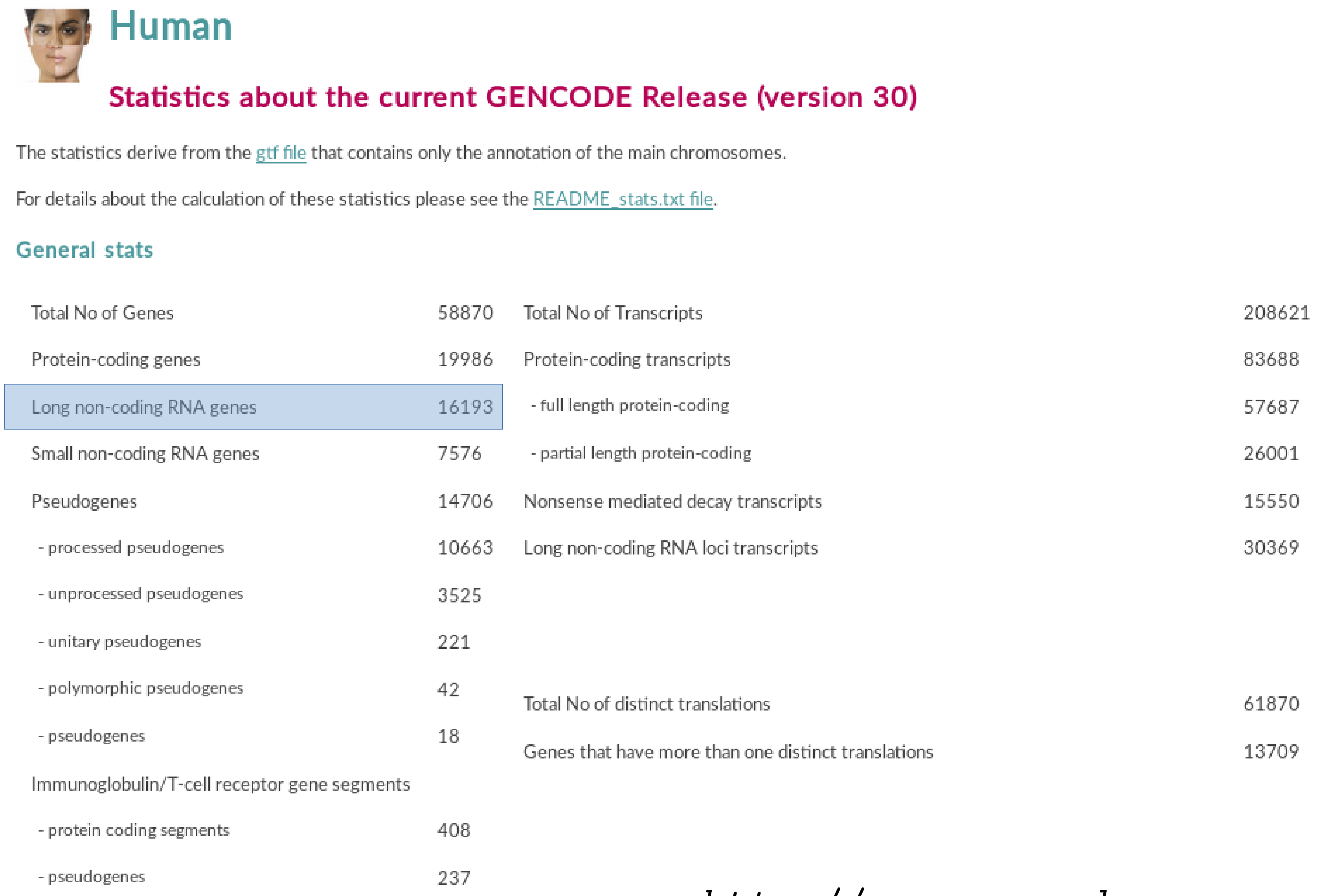
…but are "mRNA-like" transcripts:

- Transcribed by Pol_II
- Spliced
- Capped in 5’
- Most of them contain a polyA
- tail

Transcripts without coding potential , >200 nt, spliced, polyA+/- (Derrien et al., 2012)

Structural definition of lncRNAs

First Annotation in human : e.g GENCODE reference annotation
(Harrow et al., 2012, 1000 genomes project)
LncRNAs annotation has been greatly increased by the use of whole transcriptome sequencing (RNA-Seq)



Human

Statistics about the current GENCODE Release (version 30)

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

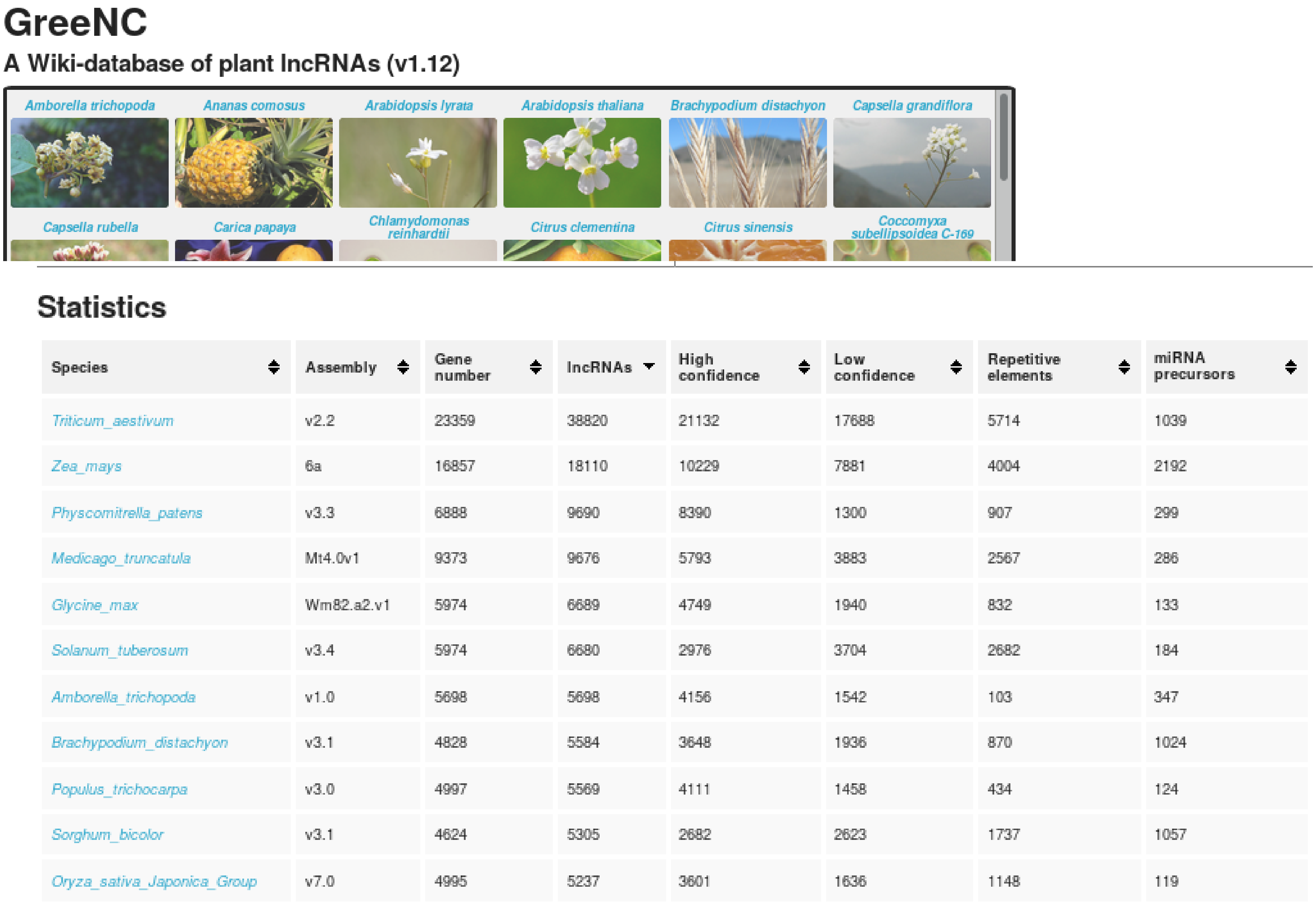
For details about the calculation of these statistics please see the README_stats.txt file.

General stats

Total No of Genes	58870	Total No of Transcripts	208621
Protein-coding genes	19986	Protein-coding transcripts	83688
Long non-coding RNA genes	16193	- full length protein-coding	57687
Small non-coding RNA genes	7576	- partial length protein-coding	26001
Pseudogenes	14706	Nonsense mediated decay transcripts	15550
- processed pseudogenes	10663	Long non-coding RNA loci transcripts	30369
- unprocessed pseudogenes	3525		
- unitary pseudogenes	221		
- polymorphic pseudogenes	42	Total No of distinct translations	61870
- pseudogenes	18	Genes that have more than one distinct translations	13709
Immunoglobulin/T-cell receptor gene segments			
- protein coding segments	408		
- pseudogenes	237		

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

Structural definition of lncRNAs



GreeNC

A Wiki-database of plant lncRNAs (v1.12)

Statistics

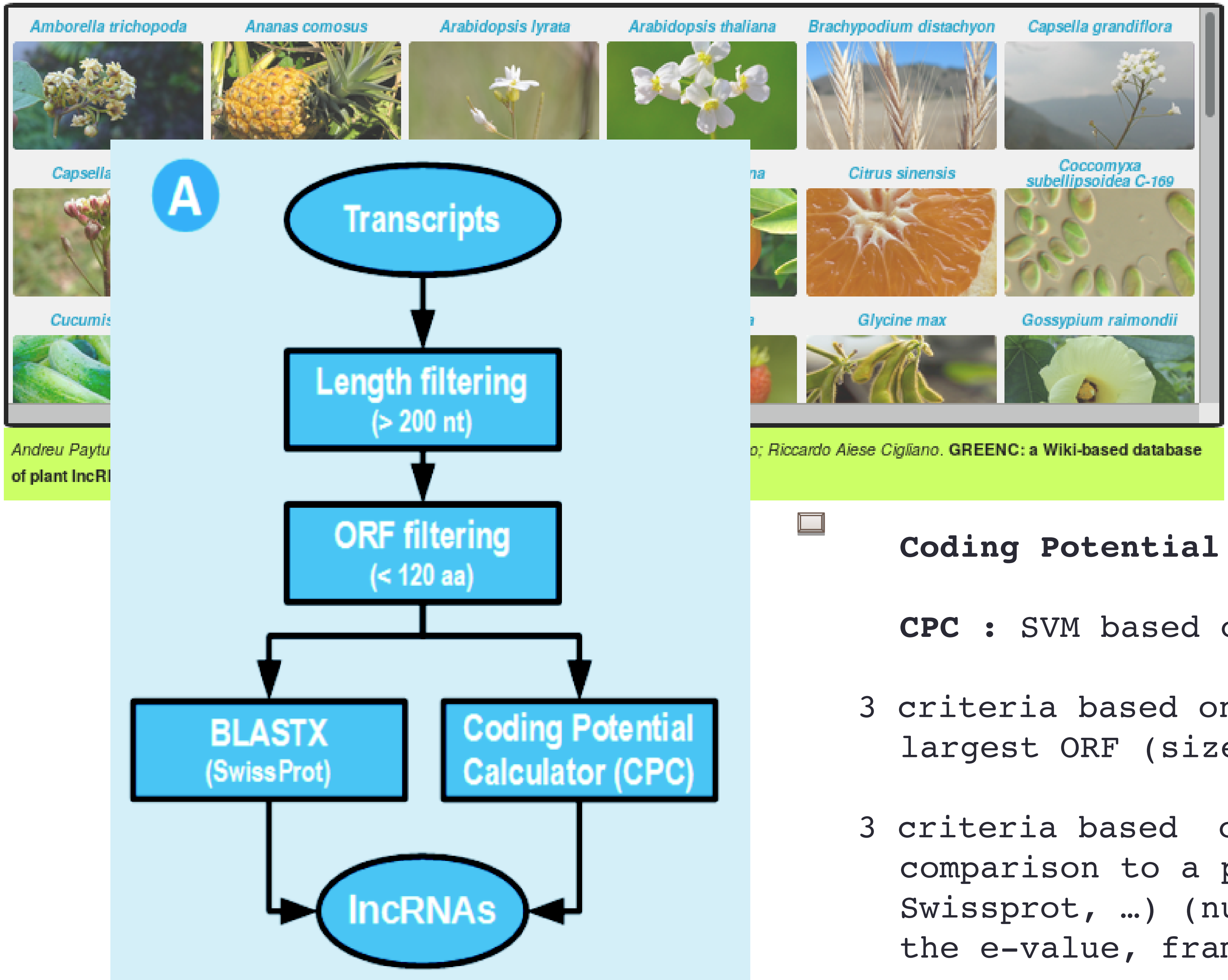
Species ⬍	Assembly ⬍	Gene number ⬍	IncRNAs ▾	High confidence ⬍	Low confidence ⬍	Repetitive elements ⬍	miRNA precursors ⬍
Triticum_aestivum	v2.2	23359	38820	21132	17688	5714	1039
Zea_mays	6a	16857	18110	10229	7881	4004	2192
Physcomitrella_patens	v3.3	6888	9690	8390	1300	907	299
Medicago_truncatula	Mt4.0v1	9373	9676	5793	3883	2567	286
Glycine_max	Wm82.a2.v1	5974	6689	4749	1940	832	133
Solanum_tuberosum	v3.4	5974	6680	2976	3704	2682	184
Amborella_trichopoda	v1.0	5698	5698	4156	1542	103	347
Brachypodium_distachyon	v3.1	4828	5584	3648	1936	870	1024
Populus_trichocarpa	v3.0	4997	5569	4111	1458	434	124
Sorghum_bicolor	v3.1	4624	5305	2682	2623	1737	1057
Oryza_sativa_Japonica_Group	v7.0	4995	5237	3601	1636	1148	119

http://greenc.sciencedesigners.com

Structural definition of lncRNAs

GreeNC

A Wiki-database of plant lncRNAs (v1.12)



Andreu Paytu … ; Riccardo Aiese Cigliano. **GREENC: a Wiki-based database of plant lncR…**

Coding Potential Calculation

CPC : SVM based on 6 parameters

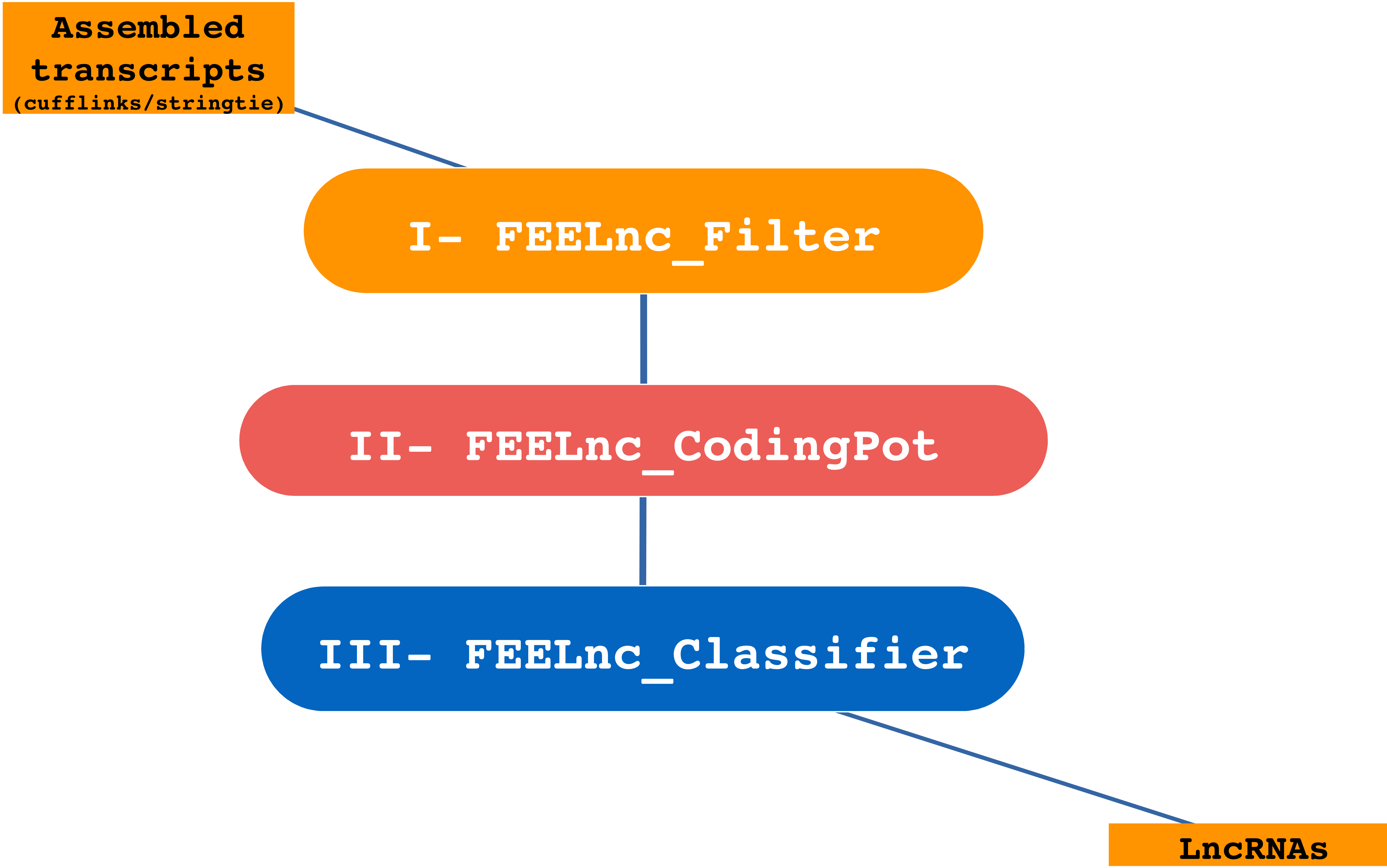
3 criteria based on the quality of the largest ORF (size, coverage, ATG)

3 criteria based on a result of a comparison to a protein databases (NR, Swissprot, …) (number of hits, mean of the e-value, frameshifts)

http://greenc.sciencedesigners.com

FEELnc : FlExible Extraction of LncRNAs

(https://github.com/tderrien/FEELnc)



Assembled transcripts (cufflinks/stringtie)

I- FEELnc_Filter

II- FEELnc_CodingPot

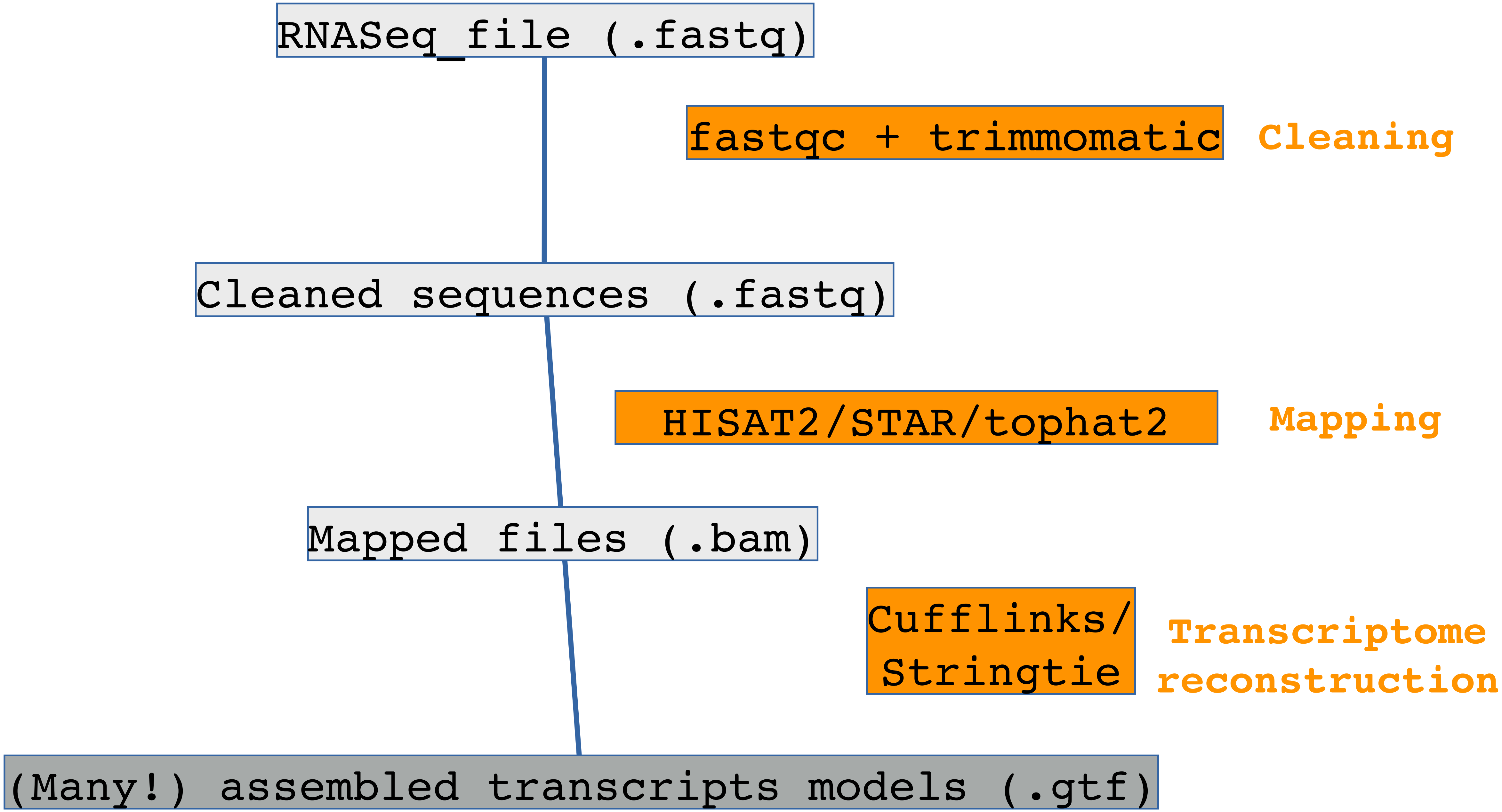
III- FEELnc_Classifier

LncRNAs

Standard pipeline for RNA-Seq analysis (genome-guided)

Input files :

- **Reference genome**
- **Reference annotation**



RNASeq_file (.fastq)

fastqc + trimmomatic

Cleaning

Cleaned sequences (.fastq)

HISAT2/STAR/tophat2

Mapping

Mapped files (.bam)

Cufflinks/
Stringtie

Transcriptome reconstruction

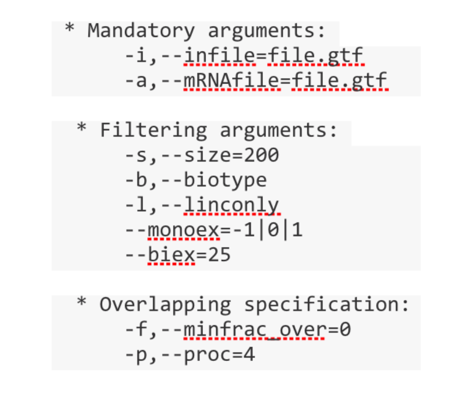
(Many!) assembled transcripts models (.gtf)

I- Filtering transcripts

Assembled transcript models (.gtf)

Known and novel transcripts

I- FEELnc_Filters



```
 * Mandatory arguments:
      -i,--infile=file.gtf
      -a,--mRNAfile=file.gtf

  * Filtering arguments:
      -s,--size=200
      -b,--biotype
      -l,--linconly
      --monoex=-1|0|1
      --biex=25

  * Overlapping specification:
      -f,--minfrac_over=0
      -p,--proc=4
```

Candidate lncRNAs

Filtering out non-lncRNAs

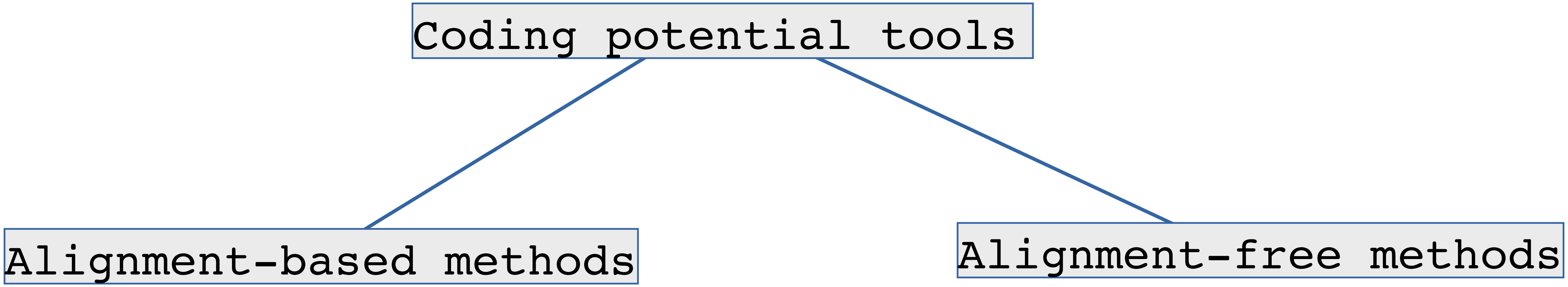
- every biotypes (small ncRNA, pseudogenes…) ?
- warning => removing potential lncRNAs host gene for small ncRNAs…
- only protein-coding biotype (=> considered as alternative isoform of mRNAs)
- < 200bp
- Monoexonic

Options :

- Fraction of overlap
- Multithreaded

II- Computing Coding Potential

Aim : define a coding potential score (CPS) for candidate transcripts and then a cutoff/threshold to differentiate mRNAs from lncRNAs

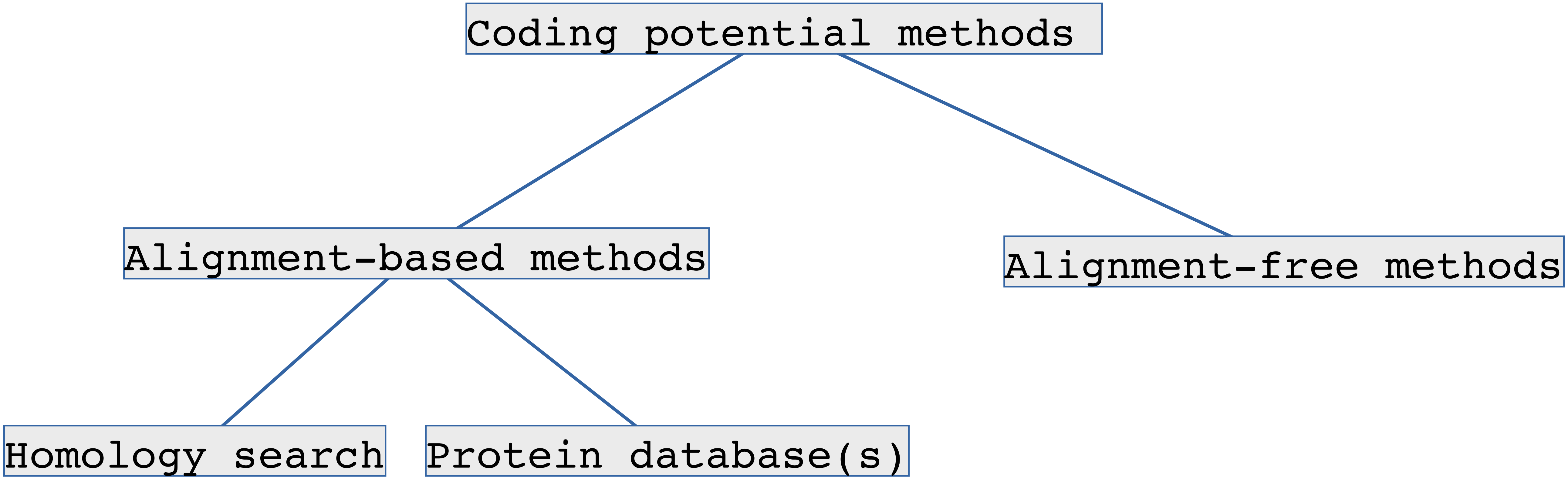


Coding potential tools

Alignment-based methods

Alignment-free methods

II- Computing Coding Potential



Coding potential methods

Alignment-based methods

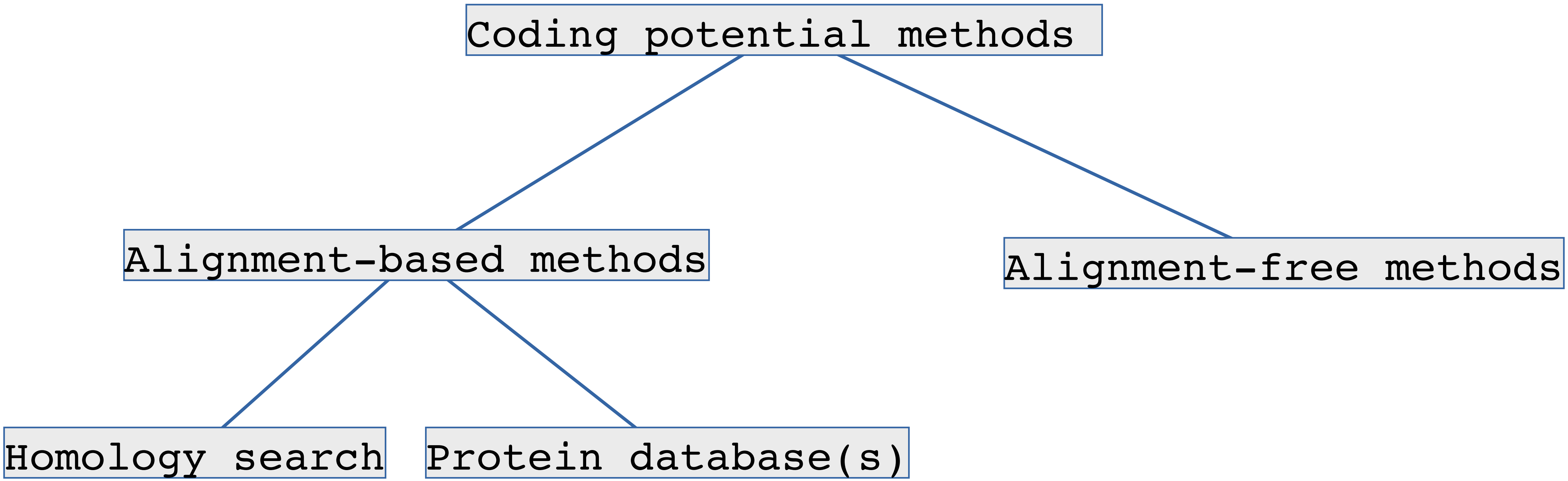
Alignment-free methods

Homology search

Protein database(s)

- **Phylo-CSF** (Lin, M. *et al*. 2011).**CPC** (Kong, L. *et al*. 2007).
- **RNACode** (Washietl, S. *et al*. 2011).

II- Computing Coding Potential



Coding potential methods

Alignment-based methods

Alignment-free methods

Homology search

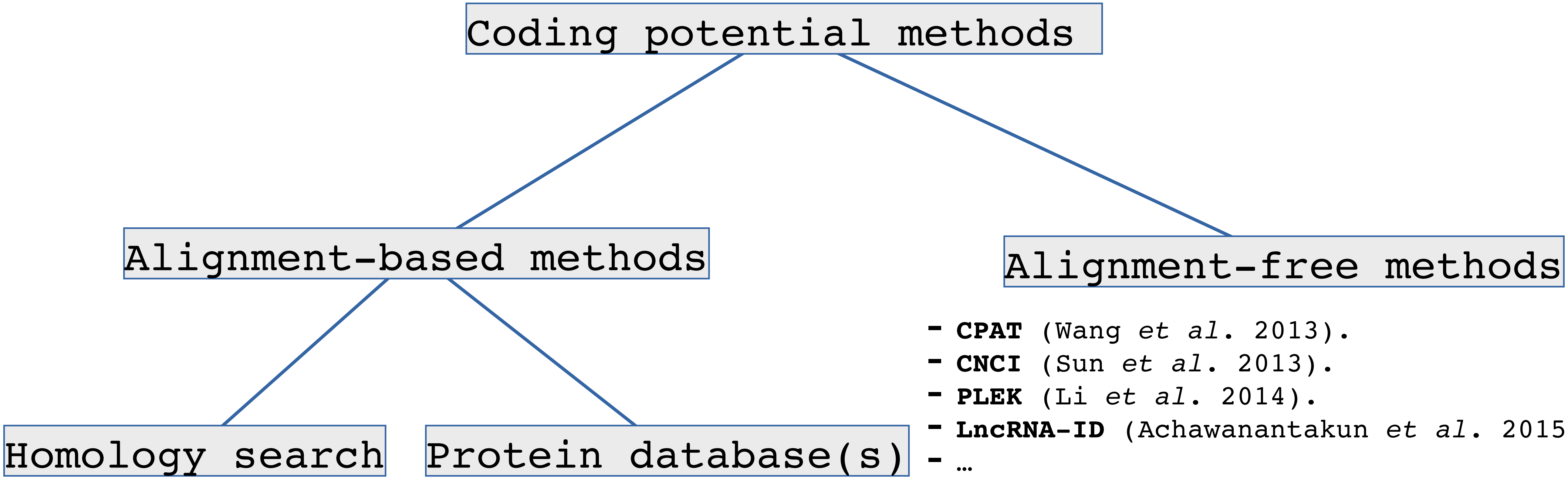
Protein database(s)

- **Phylo-CSF** (Lin, M. *et al*. 2011).**CPC** (Kong, L. *et al*. 2007).
- **RNACode** (Washietl, S. *et al*. 2011).

- Advantages:
  - high specificity
  - introduce the notion of conserved lncRNAs

- Drawbacks:
  - depends on which database/species to align with…
  - quality of the alignments
  - slow running time

II- Computing Coding Potential



- **Phylo-CSF** (Lin, M. *et al.* 2011).**CPC** (Kong, L. *et al.* 2007).
- **RNACode** (Washietl, S. *et al.* 2011).

- Advantages:
  - high specificity
  - introduce the notion of conserved lncRNAs

- Drawbacks:
  - depends on which database/species to align with…
  - quality of the alignments
  - slow running time

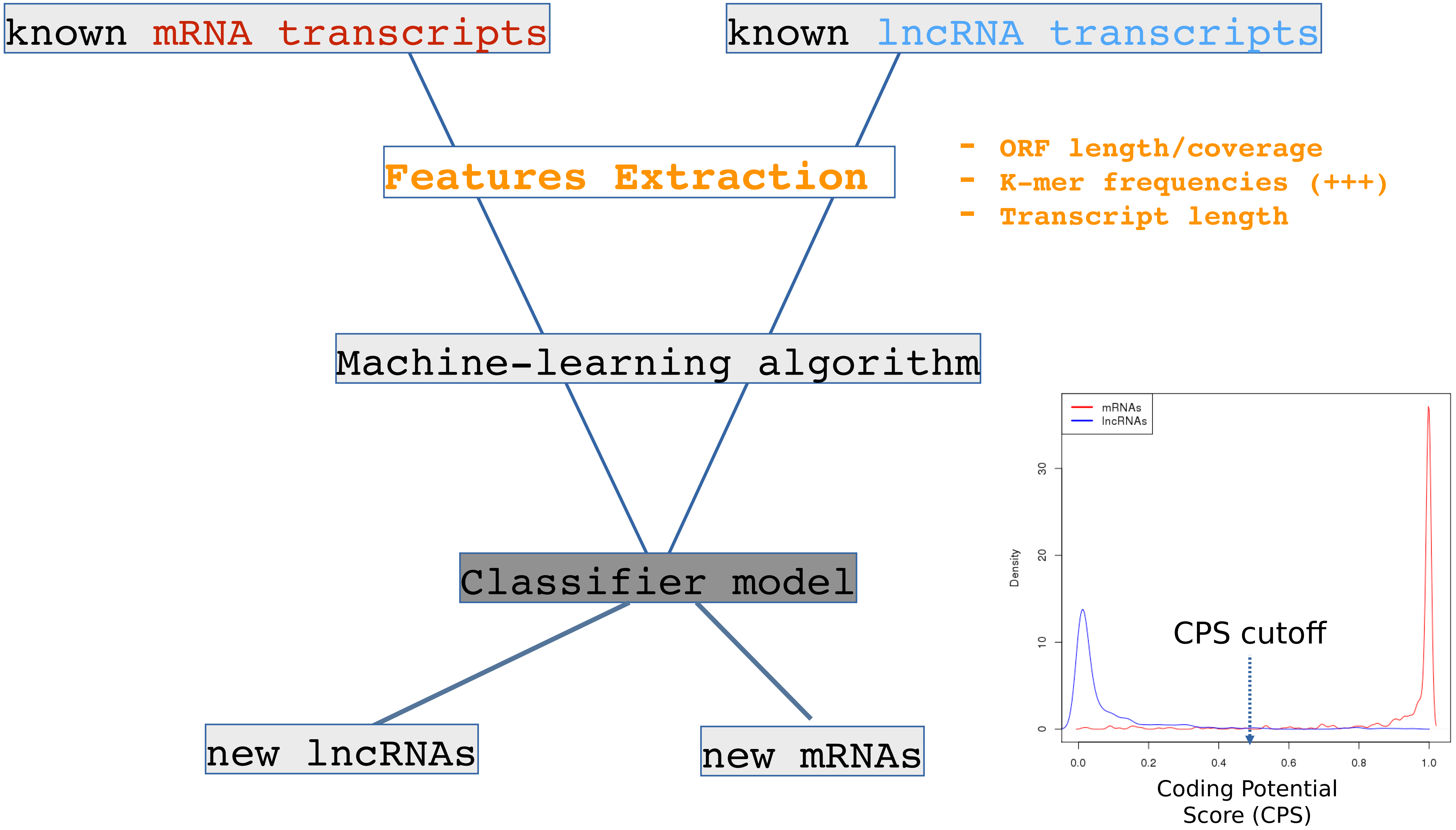
- **CPAT** (Wang *et al.* 2013).
- **CNCI** (Sun *et al.* 2013).
- **PLEK** (Li *et al.* 2014).
- **LncRNA-ID** (Achawanantakun *et al.* 2015).
- …
- **COME** (Hu *et al.* 2016).
- **lncScore** (Zhao *et al.* 2016).

- Advantages:
  - fast and (relatively) easy to use
  - independent of any alignment
  - lineage-specific lncRNAs (or weakly conserved)

- Drawbacks:
  - +++ designed for model organisms

II- Computing Coding Potential

Extracting Features and Machine learning



known mRNA transcripts

known lncRNA transcripts

Features Extraction

- ORF length/coverage
- K-mer frequencies (+++)
- Transcript length

Machine-learning algorithm

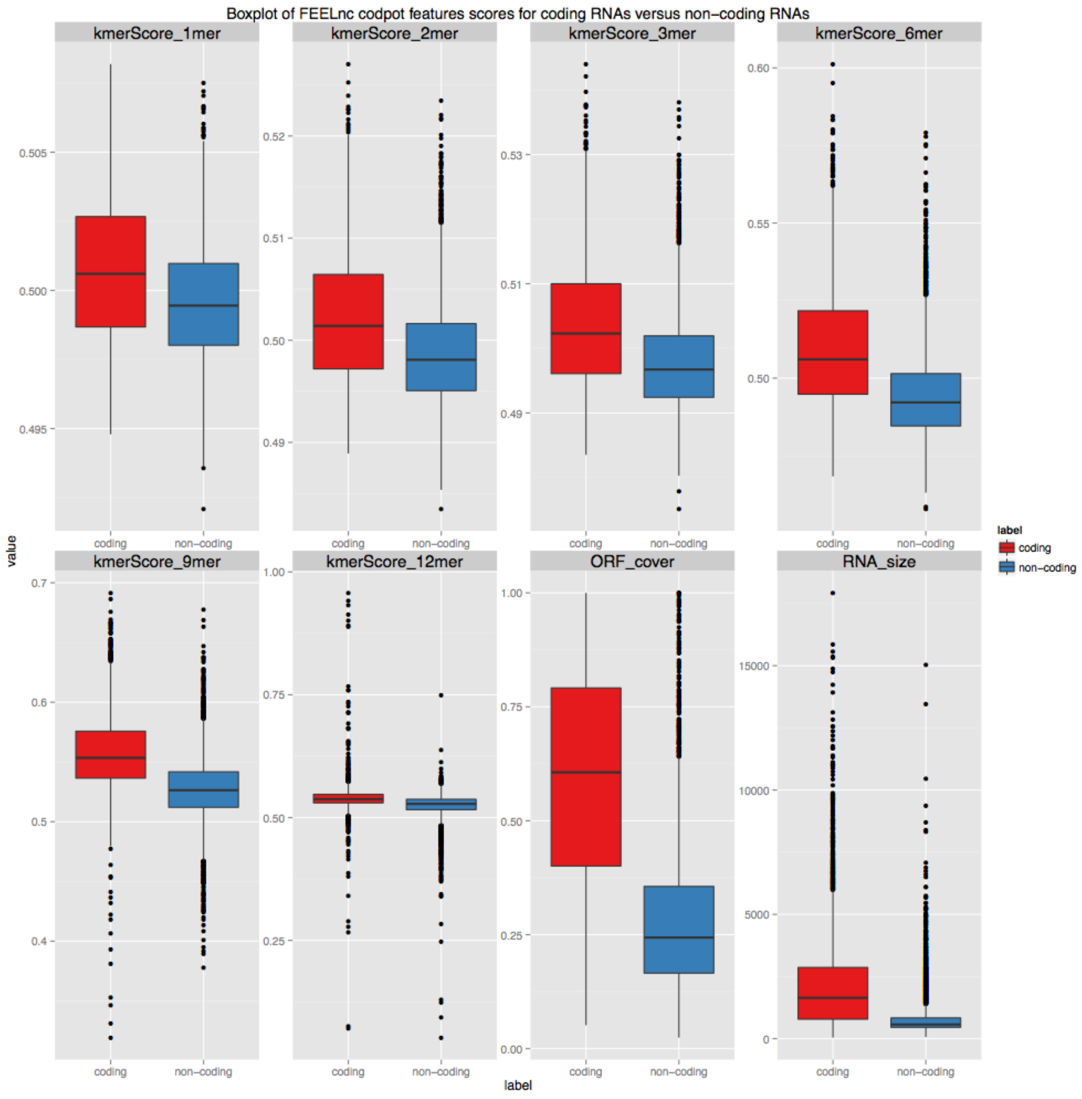
Classifier model

new lncRNAs

new mRNAs

3 main classifiers/features

1. Multi-*k*-mer scores (*k= 1 to 12*)
   - **KmerInShort (KIS)** developed from GATB tools (Drezen et al.) (https://github.com/rizkg/KmerInShort)
   - very fast and parallel extraction of k-mer profiles (not limited to one k-mer i. hexamer)
2. ORF coverage (ORF defined wrt 3 modes)
   - strict : requires start && stop
   - moderate : requires start || stop
   - relaxed : RNA sequence
3. RNA size



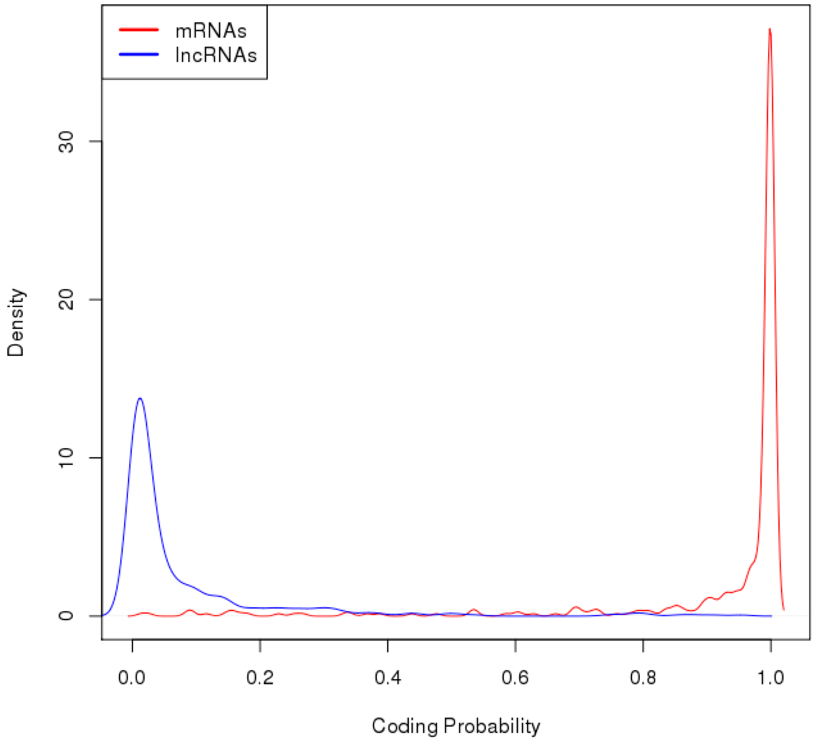
II- Computing Coding Potential

Feature scores on known lncRNAs and mRNAs

RandomForest

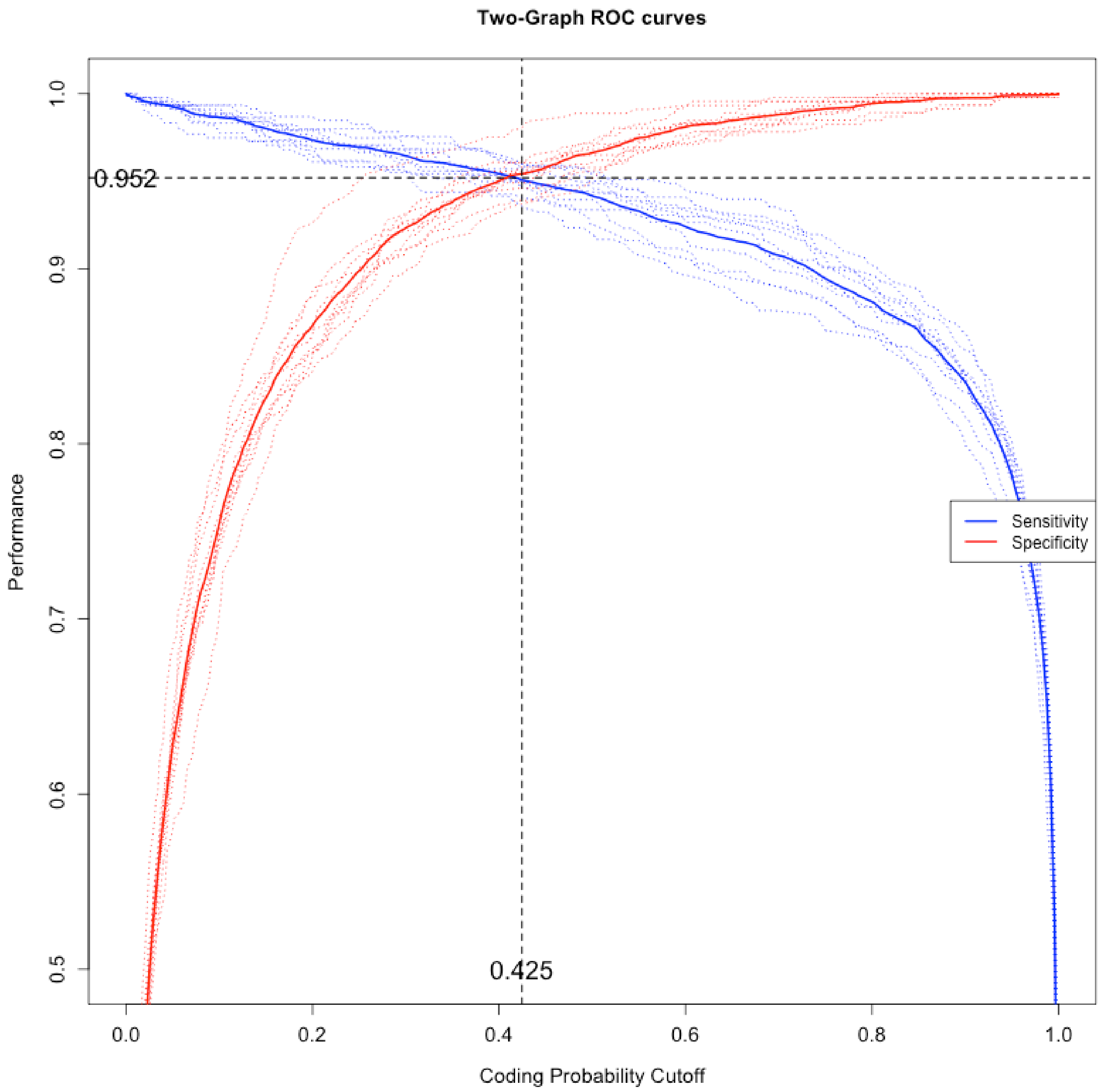
- fast and easy to optimize
- can deal with unbalanced training set (+++)
- could deal with missing data

Defines a coding potential score (CPS) for all transcripts from training sets



What is the best CPS cutoff?

Optimal CPS threshold



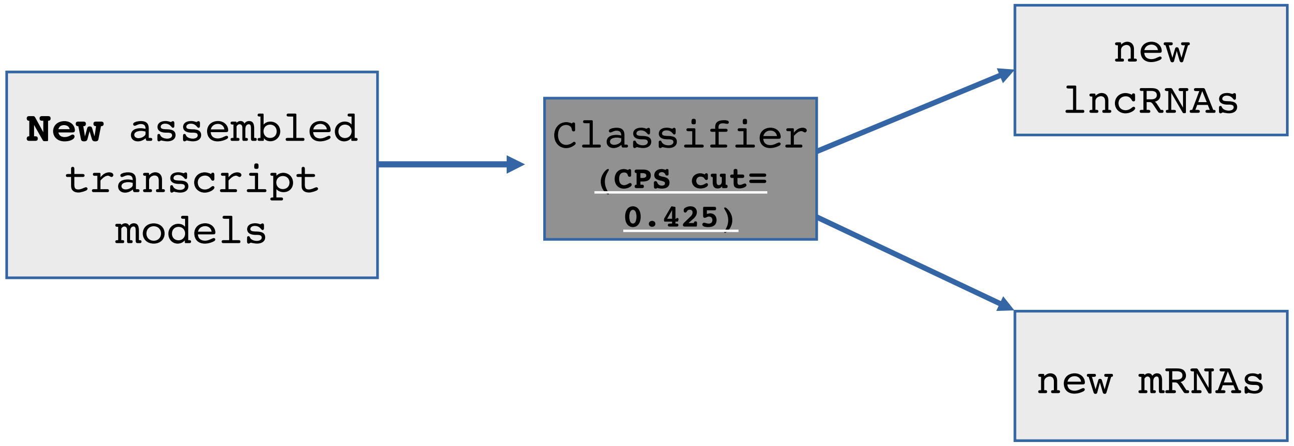
Coding Potential Score (CPS)

- Automatic Two-Graph ROC curves to define an optimal cutoff to separate mRNAs from lncRNAs
- FEELnc defined optimal CPS cutoff via :
-

Sensitivity == Specificity (= 0.952)

or

Max (Sn & Sp)



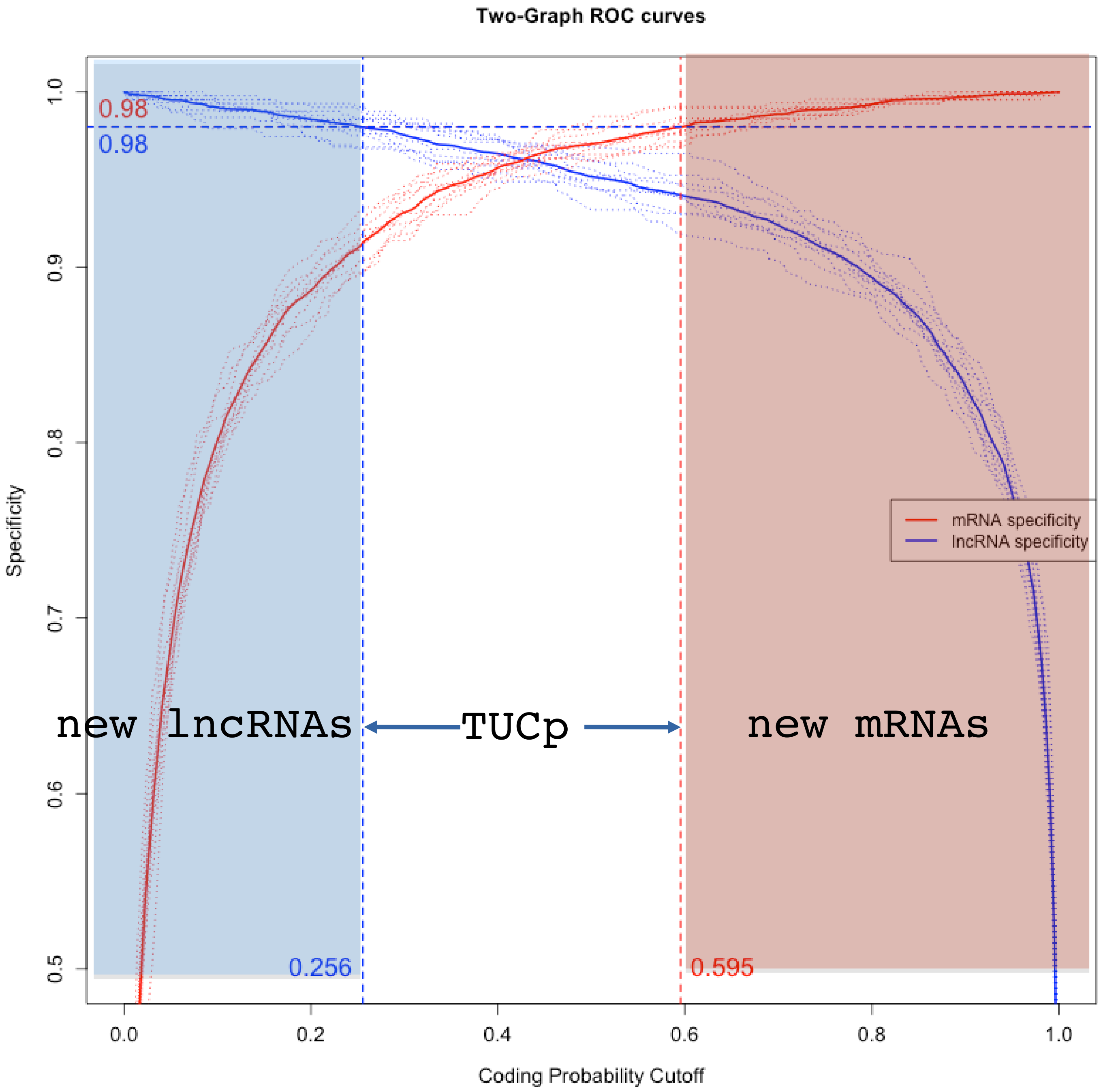
New assembled transcript models

Classifier **(CPS cut= 0.425)**

new lncRNAs

new mRNAs

Optimal CPS threshold

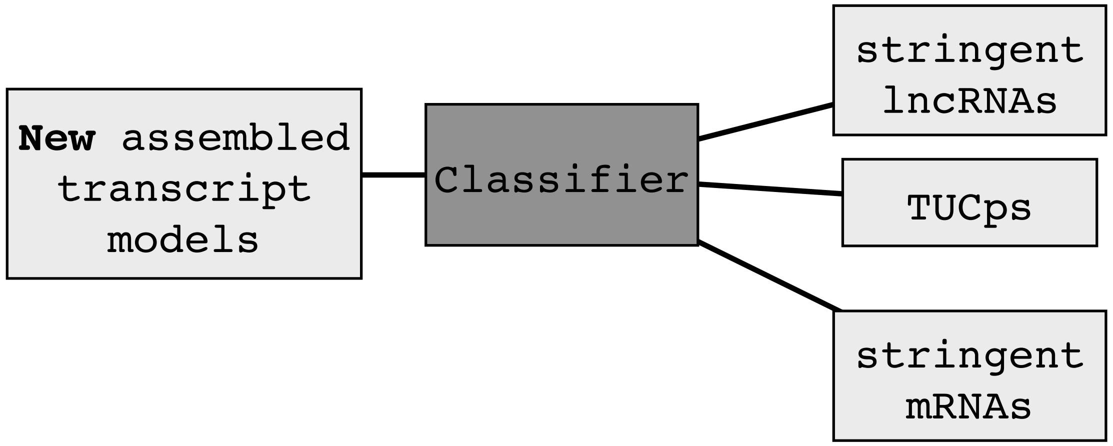


Coding Potential Score (CPS)

"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.

- TUCPs = ambiguous transcript given their CPS



New assembled transcript models

Classifier

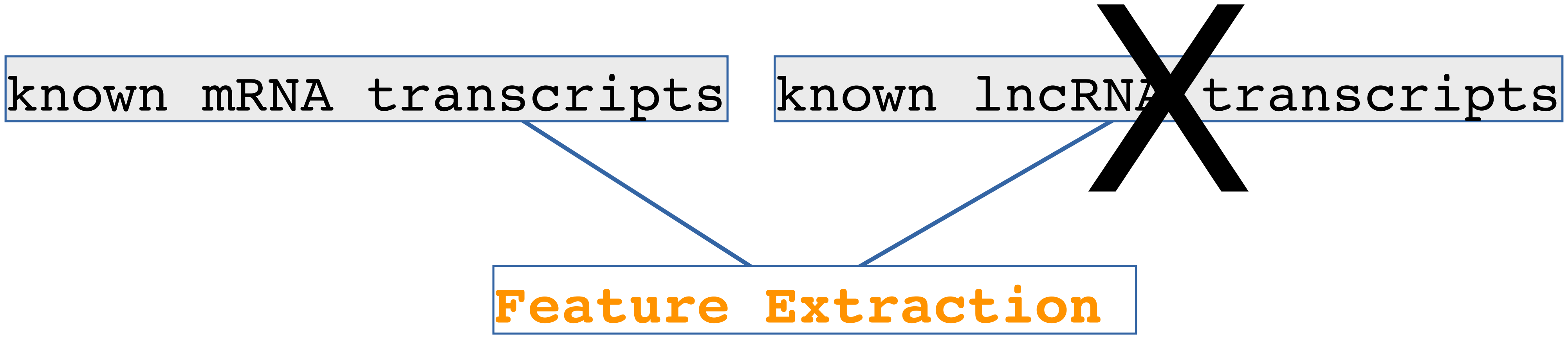
stringent lncRNAs

TUCps

stringent mRNAs

FEELnc : without lncRNA training set

Training



known mRNA transcripts

known lncRNA transcripts

X

Feature Extraction

For non-model organisms, known lncRNAs are (often) not available.
FEELnc implements 2 options to model ncRNAs:

- **intergenic module :** extract random intergenic sequences
- **mRNA shuffle module :** shuffle mRNA sequences while preserving a certain *k*-mer frequency (using UShuffle program)

II- Computing Coding Potential

Tools performance on the GENCODE human datasets.

F-score/MCC capture the global performance
of the tools in a single measure

HUMAN dataset	Program	Sensitivity	Specificity	Precision	Accuracy	F-score	MCC
CPC	**FEELnc**	**0.923**	0.915	0.916	**0.919**	**0.919**	**0.838**
	CPAT	0.899	0.924	0.922	0.912	0.910	0.823
	CPAT_train	0.920	0.901	0.903	0.910	0.911	0.821
	CNCI	0.829	0.979	0.975	0.904	0.896	0.817
	PLEK	0.732	**0.985**	**0.981**	0.858	0.838	0.741
	PhyloCSF	0.906	0.802	0.820	0.854	0.861	0.712
	PLEK_train	0.582	0.960	0.936	0.770	0.718	0.584
	0.728	0.719	0.713	0.438			

Bold-underlined values correspond to the highest values of each metrics.

CPAT_train and PLEK_train correspond to program versions trained with the human training dataset.

Programs are sorted by MCC values.

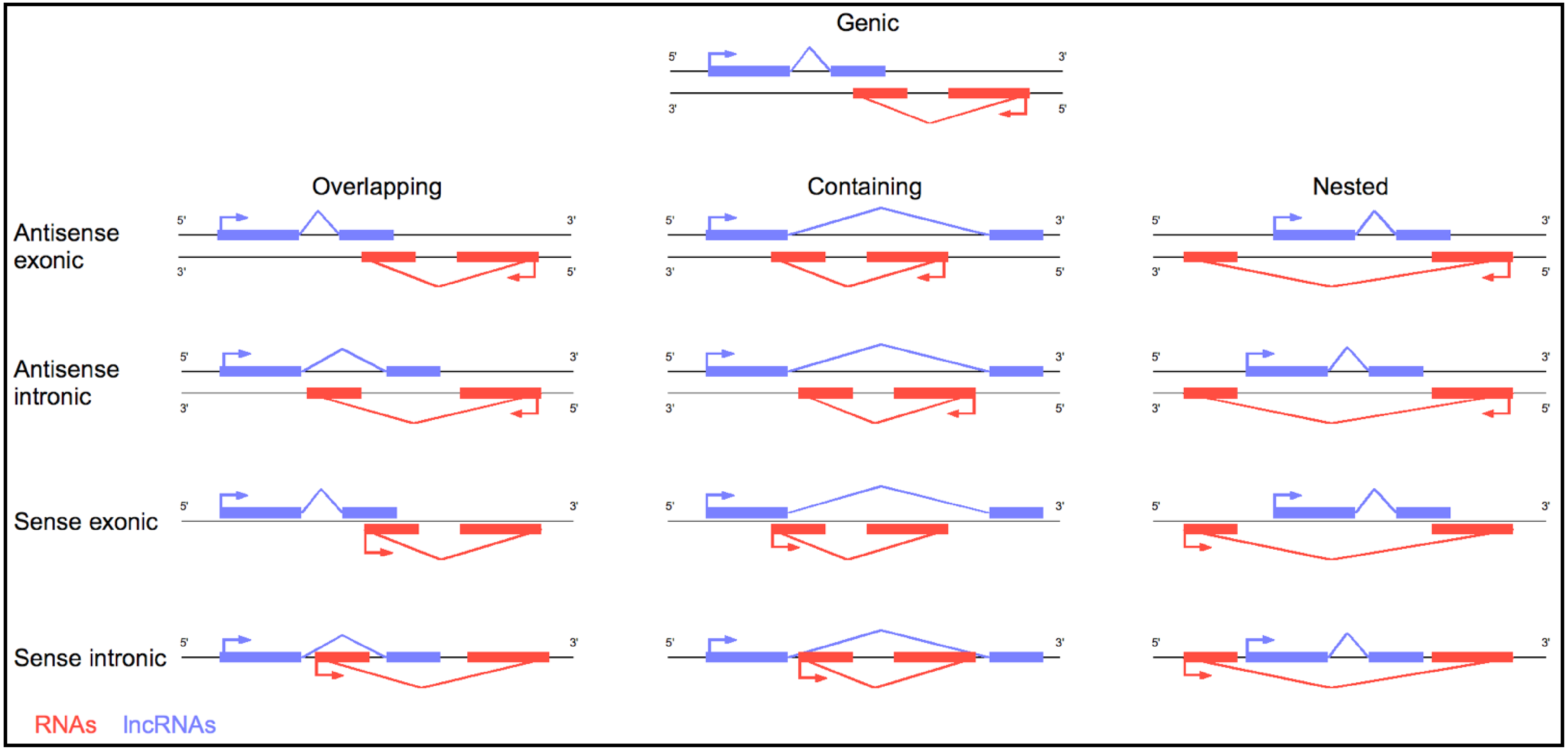
III- Classifying

- Classifying lncRNAs genomic context wrt to mRNAs could help predict functionality
- FEELnc_Classifier uses a sliding widow around lncRNA
- Classify with closest RNA_partner (mRNAs or ncRNAs) according to
  - orientation of transcription
  - location of the interaction

- FEELnc classifier output file:

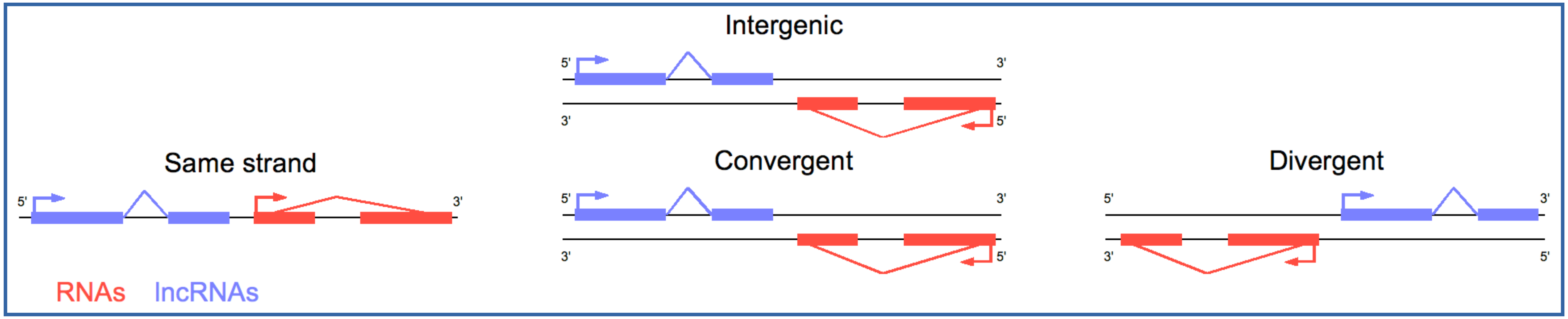
```
cat{INPUT}_classes.txt
isBest lncRNA_gene lncRNA_transcript partnerRNA_gene   partnerRNA_transcript direction type     distance subtype    location
1    XLOC_090743 TCONS_00232056   ENSCAFG00000013346 ENSCAFT00000021186   antisense intergenic 377     divergent  upstream
1    XLOC_090720 TCONS_00231943   ENSCAFG00000026373 ENSCAFT00000040656   sense    intergenic 66670   same_strand upstream
1    XLOC_090678 TCONS_00231794   ENSCAFG00000010781 ENSCAFT00000017151   antisense genic    0      nested     intronic
0    XLOC_090678 TCONS_00231794   ENSCAFG00000010794 ENSCAFT00000017171   sense    intergenic 8293    same_strand upstream
```

III- Classifying



Such as gencode classes (i.e sense_intronic,antisense…)

III- Classifying



dubious lincRNAs
(i.e UTRs of mRNAs)

lincRNAs sharing bi-directional promoter with mRNAs

Availability

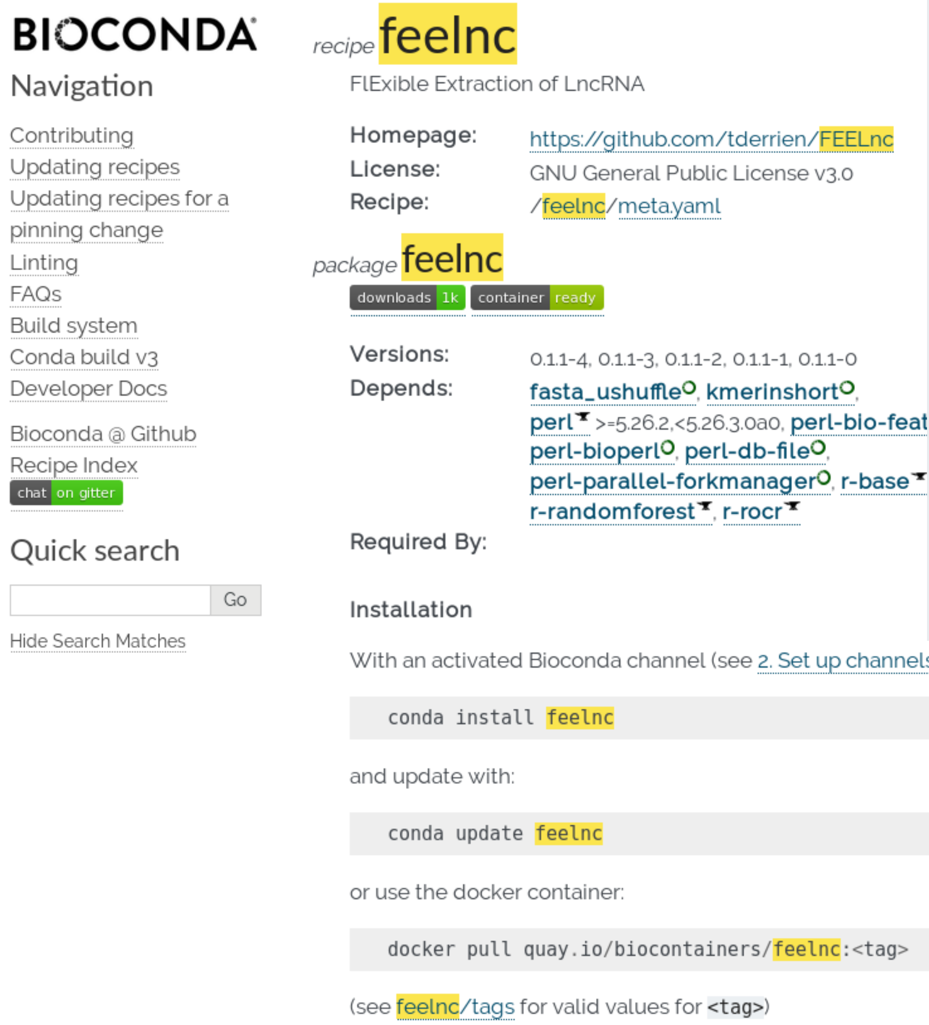
- Available on **Github:**

  https://github.com/tderrien/FEELnc

- A recipe **Bioconda**
- A **nextflow/docker** implementation of STAR/Cufflink/FEELnc pipeline has been developed by Evan FLODEN :

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

- **Galaxy toolshed**



recipe feelnc

FlExible Extraction of LncRNA

Homepage: https://github.com/tderrien/FEELnc
License: GNU General Public License v3.0
Recipe: /feelnc/meta.yaml

package feelnc

downloads 1k container ready

Versions: 0.1.1-4, 0.1.1-3, 0.1.1-2, 0.1.1-1, 0.1.1-0
Depends: fasta_ushuffle, kmerinshort, perl >=5.26.2,<5.26.3.0a0, perl-bio-feat perl-bioperl, perl-db-file, perl-parallel-forkmanager, r-base r-randomforest, r-rocr

Required By:

Installation

With an activated Bioconda channel (see 2. Set up channels), install with:

```
conda install feelnc
```

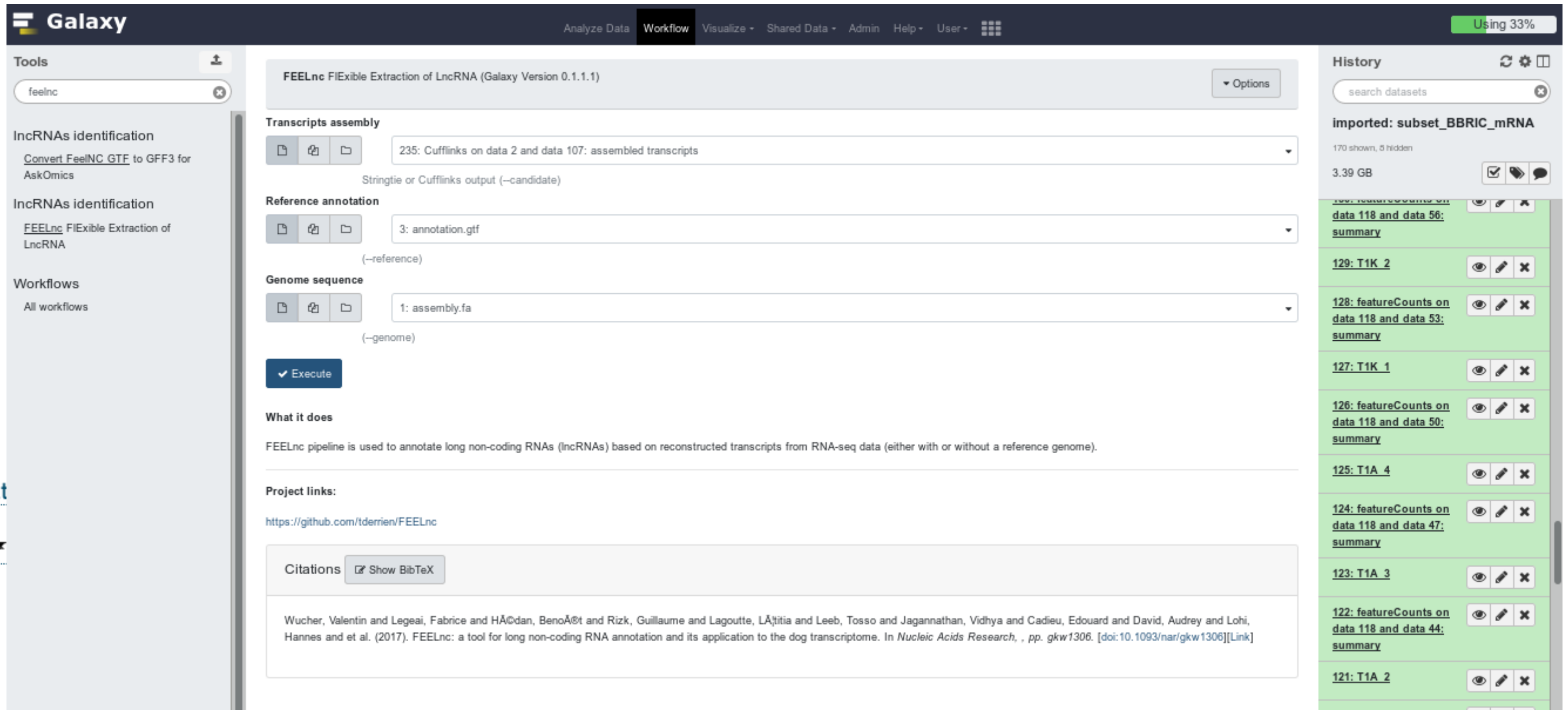
and update with:

```
conda update feelnc
```

or use the docker container:

```
docker pull quay.io/biocontainers/feelnc:<tag>
```

(see feelnc/tags for valid values for `<tag>`)



FEELnc FlExible Extraction of LncRNA (Galaxy Version 0.1.1.1)

Transcripts assembly: 235: Cufflinks on data 2 and data 107: assembled transcripts
Stringtie or Cufflinks output (--candidate)

Reference annotation: 3: annotation.gtf
(--reference)

Genome sequence: 1: assembly.fa
(--genome)

What it does

FEELnc pipeline is used to annotate long non-coding RNAs (lncRNAs) based on reconstructed transcripts from RNA-seq data (either with or without a reference genome).

Project links:

https://github.com/tderrien/FEELnc

Citations

Wucher, Valentin and Legeai, Fabrice and HÃ©dan, BenoÃ®t and Rizk, Guillaume and Lagoutte, LÃ¦titia and Leeb, Tosso and Jagannathan, Vidhya and Cadieu, Edouard and David, Audrey and Lohi, Hannes and et al. (2017). FEELnc: a tool for long non-coding RNA annotation and its application to the dog transcriptome. In *Nucleic Acids Research, , pp. gkw1306.* [doi:10.1093/nar/gkw1306][Link]

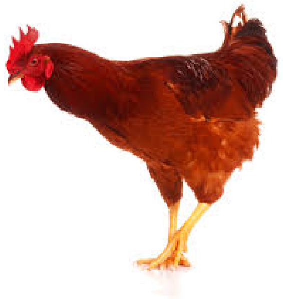
imported: subset_BBRIC_mRNA
170 shown, 5 hidden
3.39 GB

data 118 and data 56: summary
129: T1K_2
128: featureCounts on data 118 and data 53: summary
127: T1K_1
126: featureCounts on data 118 and data 50: summary
125: T1A_4
124: featureCounts on data 118 and data 47: summary
123: T1A_3
122: featureCounts on data 118 and data 44: summary
121: T1A_2

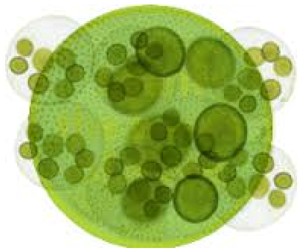
https://galaxy.genouest.org/

Applications

- Various tissues (n=16 in 7 different breeds)
- ~2,500 new lncRNAs loci (lincRNAs)
- Wucher V. *et al*, NAR 2017



- Adipose and liver tissues
- 2,200 novel lncRNA genes
- Muret K. *et al*, Genet Sel Evol 2017

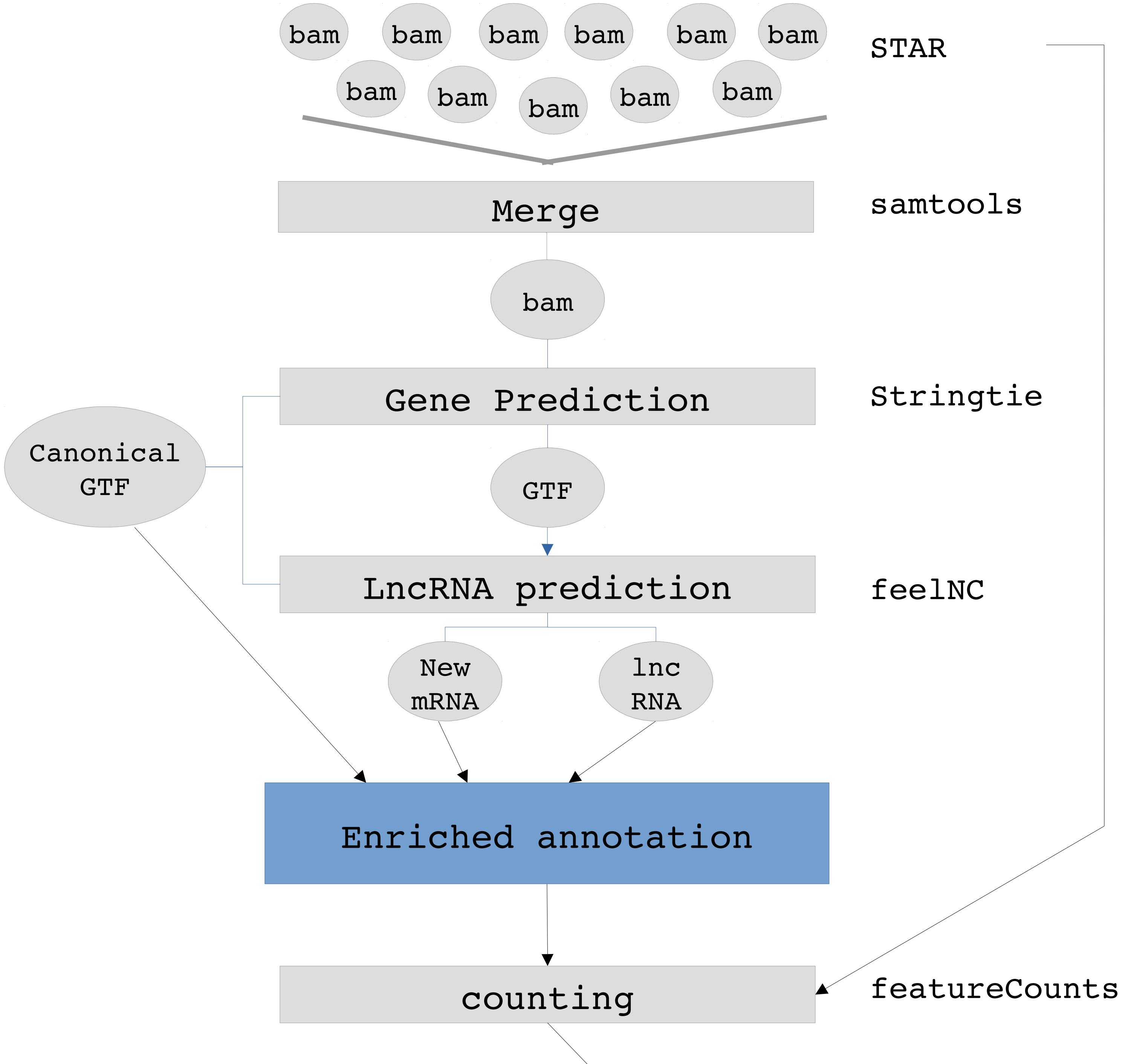


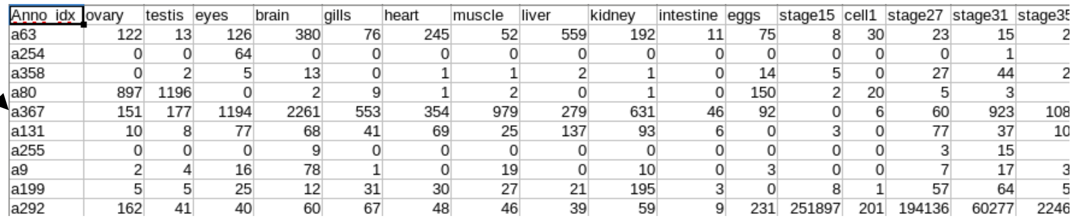
- male and femelle gametophytes
- 717 novel lncRNA genes
- Cormier A. *et al*, New Phytol 2017

- Sexual and asexual embryos ; salivary glands of plant host biotype
- 2625 novel lncRNA genes
- In preparation

- Ovogenesis medaka
- 1131 novel lncRNA genes
- In preparation

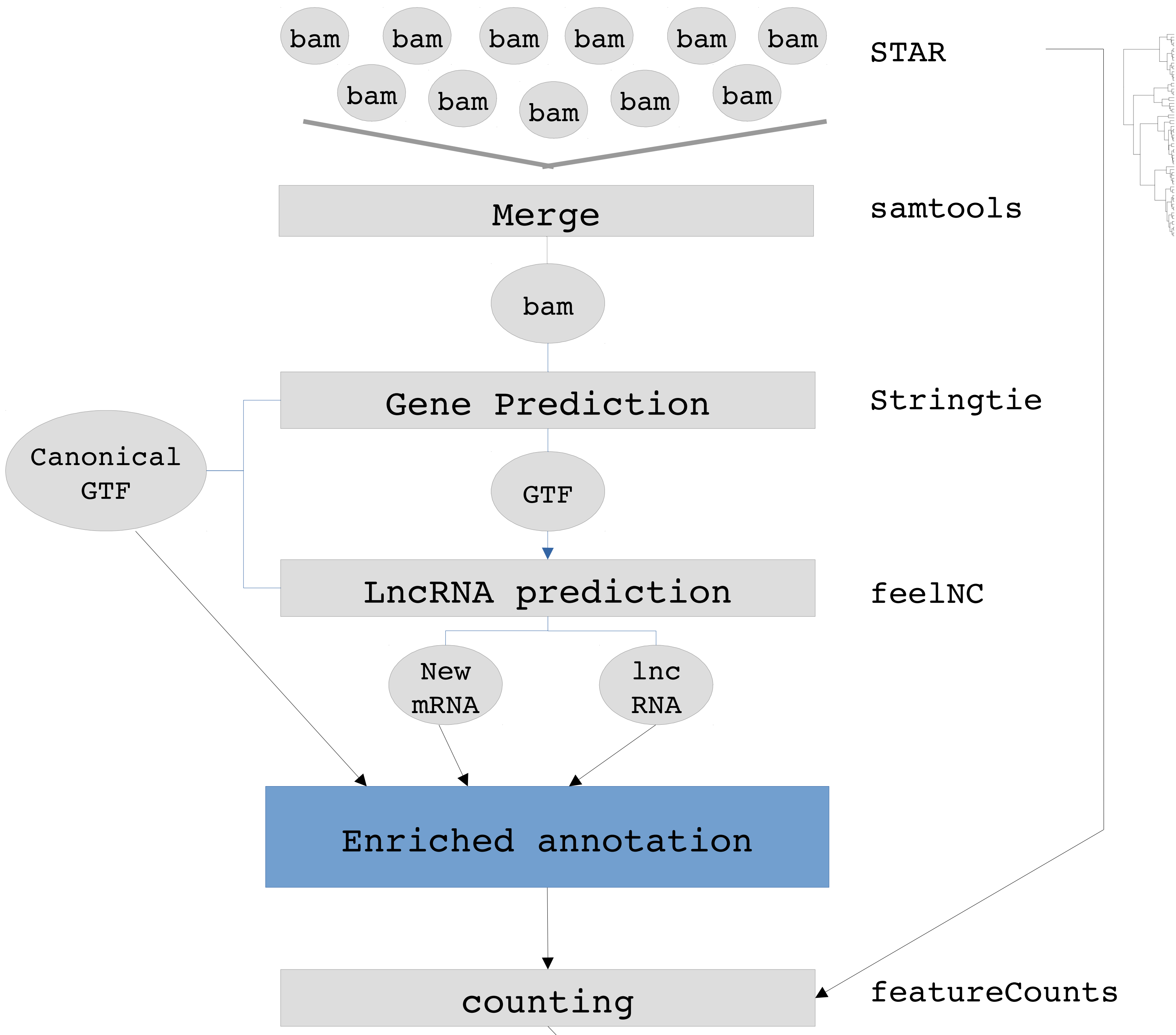
Applications

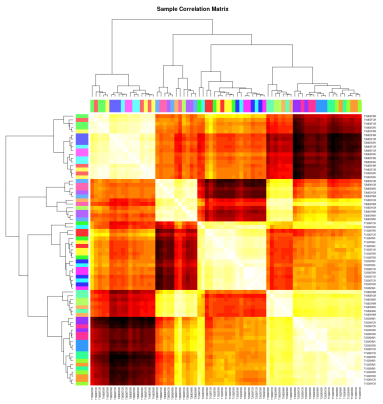


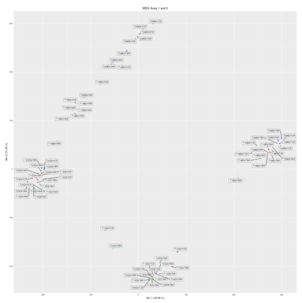


Anno_idx	ovary	testis	eyes	brain	gills	heart	muscle	liver	kidney	intestine	eggs	stage15	cell1	stage27	stage31	stage35
a63	122	13	126	380	76	245	52	559	192	11	75	8	30	23	15	2
a254	0	0	64	0	0	0	0	0	0	0	0	0	0	0	1	
a358	0	2	5	13	0	1	1	2	1	0	14	5	0	27	44	2
a80	897	1196	0	2	9	1	2	0	1	0	150	2	20	5	3	
a367	151	177	1194	2261	553	354	979	279	631	46	92	0	6	60	923	108
a131	10	8	77	68	41	69	25	137	93	6	0	3	0	77	37	10
a255	0	0	0	9	0	0	0	0	0	0	0	0	0	3	15	
a9	2	4	16	78	1	0	19	0	10	0	3	0	0	7	17	3
a199	5	5	25	12	31	30	27	21	195	3	0	8	1	57	64	5
a292	162	41	40	60	67	48	46	39	59	9	231	251897	201	194136	60277	2246

Applications

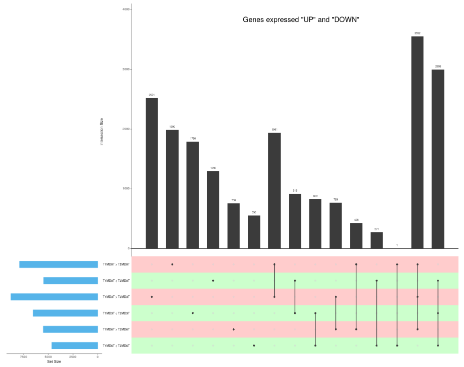


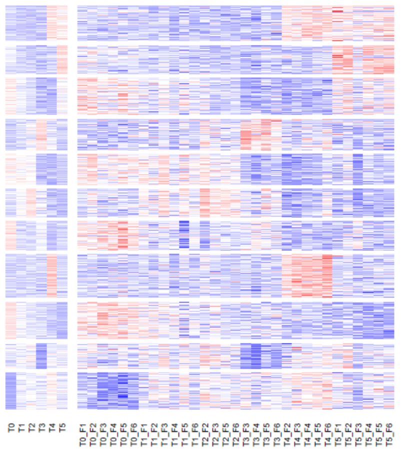


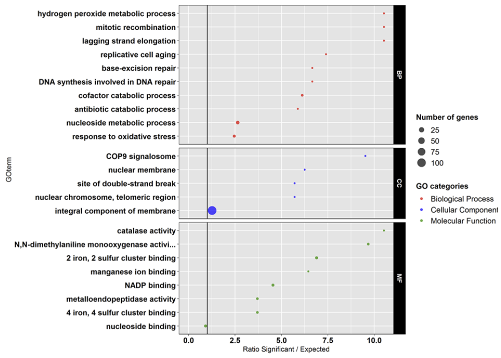


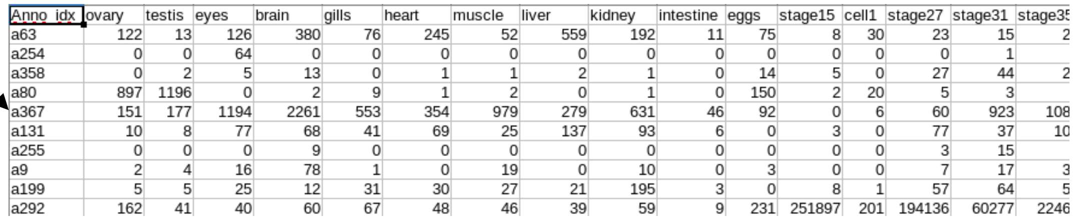
```
Differentialexpressions analysis


T1SD0T<T1SD3T T1SD0T=T1SD3T T1SD0T>T1SD3T
      15      64633       189
T1SD0T<T1SD6T T1SD0T=T1SD6T T1SD0T>T1SD6T
     1180      63479       178
T1SD3T<T1SD6T T1SD3T=T1SD6T T1SD3T>T1SD6T
     1002      63825        10
T2SD0T<T2SD3T T2SD0T=T2SD3T T2SD0T>T2SD3T
      440      63225       1172
T2SD0T<T2SD6T T2SD0T=T2SD6T T2SD0T>T2SD6T
     1767      59938       3132
T2SD3T=T2SD6T
    64837
T1SD0Y=T1SD3Y
    64837
T1SD0Y<T1SD6Y T1SD0Y=T1SD6Y
       5     64832
```

Asko





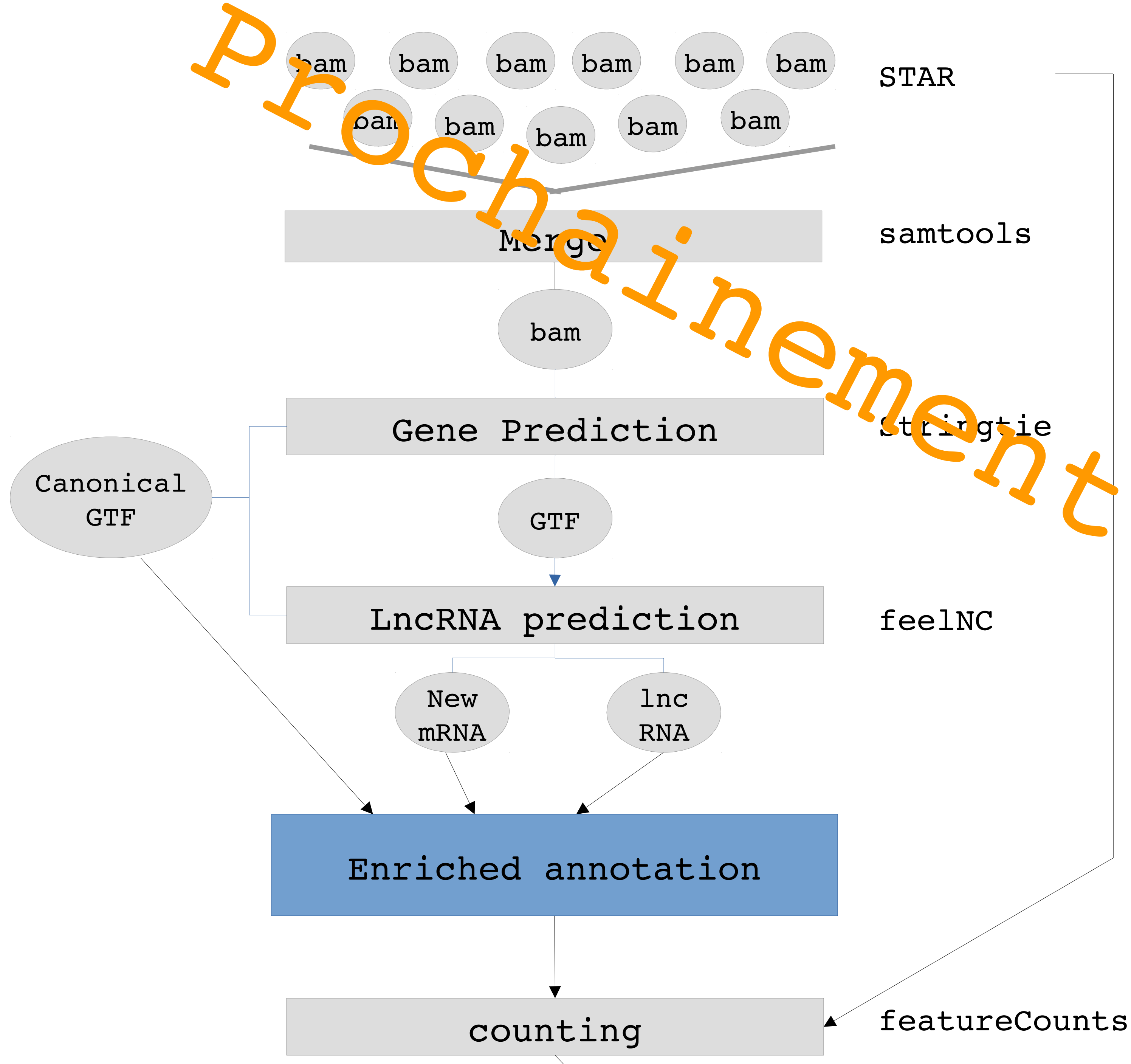




Anno_idx	ovary	testis	eyes	brain	gills	heart	muscle	liver	kidney	intestine	eggs	stage15	cell1	stage27	stage31	stage35
a63	122	13	126	380	76	245	52	559	192	11	75	8	30	23	15	2
a254	0	0	64	0	0	0	0	0	0	0	0	0	0	0	1	
a358	0	2	5	13	0	1	1	2	1	0	14	5	0	27	44	2
a80	897	1196	0	2	9	1	2	0	1	0	150	2	20	5	3	
a367	151	177	1194	2261	553	354	979	279	631	46	92	0	6	60	923	108
a131	10	8	77	68	41	69	25	137	93	6	0	3	0	77	37	10
a255	0	0	0	9	0	0	0	0	0	0	0	0	0	3	15	
a9	2	4	16	78	1	0	19	0	10	0	3	0	0	7	17	3
a199	5	5	25	12	31	30	27	21	195	3	0	8	1	57	64	5
a292	162	41	40	60	67	48	46	39	59	9	231	251897	201	194136	60277	2246

Applications

Prochainement en Nextflow



bam bam bam bam bam bam

bam bam bam bam bam

STAR

Merge

samtools

bam

Gene Prediction

stringtie

Canonical GTF

GTF

LncRNA prediction

feelNC

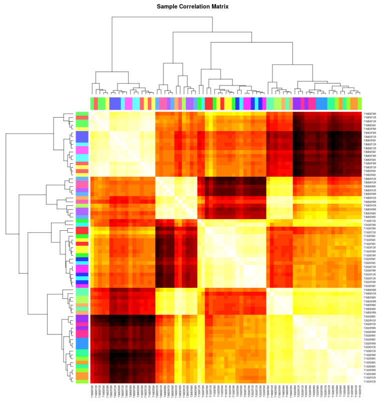
New mRNA

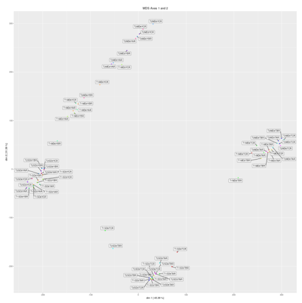
lnc RNA

Enriched annotation

counting

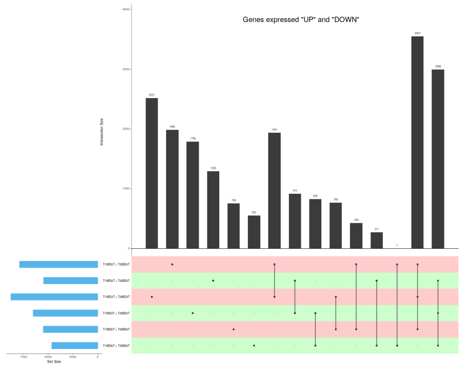
featureCounts

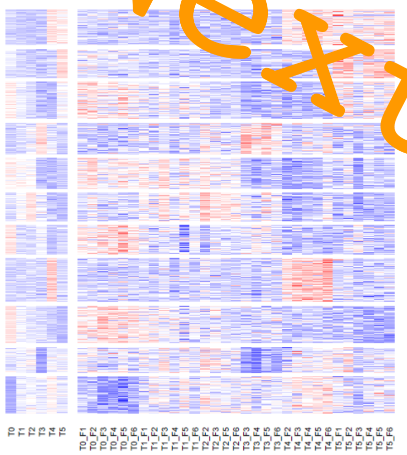


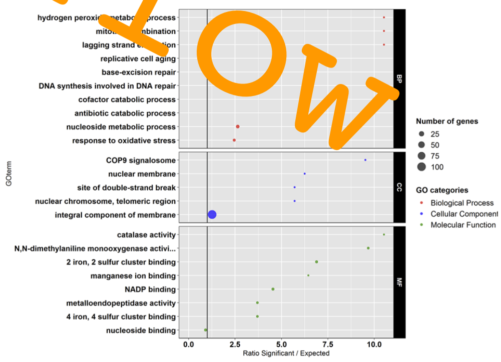


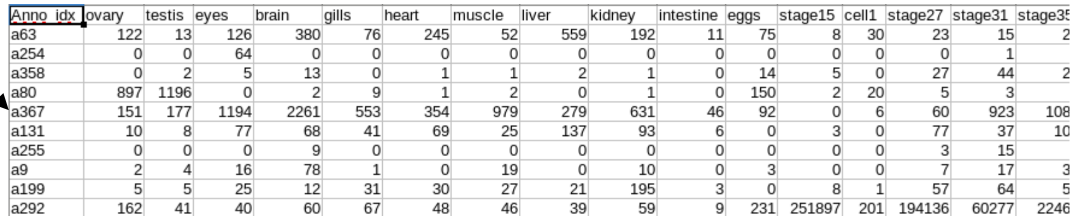
```
Differentialexpressions analysis


T1SD0T<T1SD3T T1SD0T=T1SD3T T1SD0T>T1SD3T
      15      64633       189
T1SD0T<T1SD6T T1SD0T=T1SD6T T1SD0T>T1SD6T
    1180      63479       178
T1SD3T<T1SD6T T1SD3T=T1SD6T T1SD3T>T1SD6T
    1002      63825        10
T2SD0T<T2SD3T T2SD0T=T2SD3T T2SD0T>T2SD3T
     440      63225      1172
T2SD0T<T2SD6T T2SD0T=T2SD6T T2SD0T>T2SD6T
    1767      59938      3132
T2SD3T=T2SD6T
    64837
T1SD0Y=T1SD3Y
    64837
T1SD0Y<T1SD6Y T1SD0Y=T1SD6Y
       5      64832
```

Asko







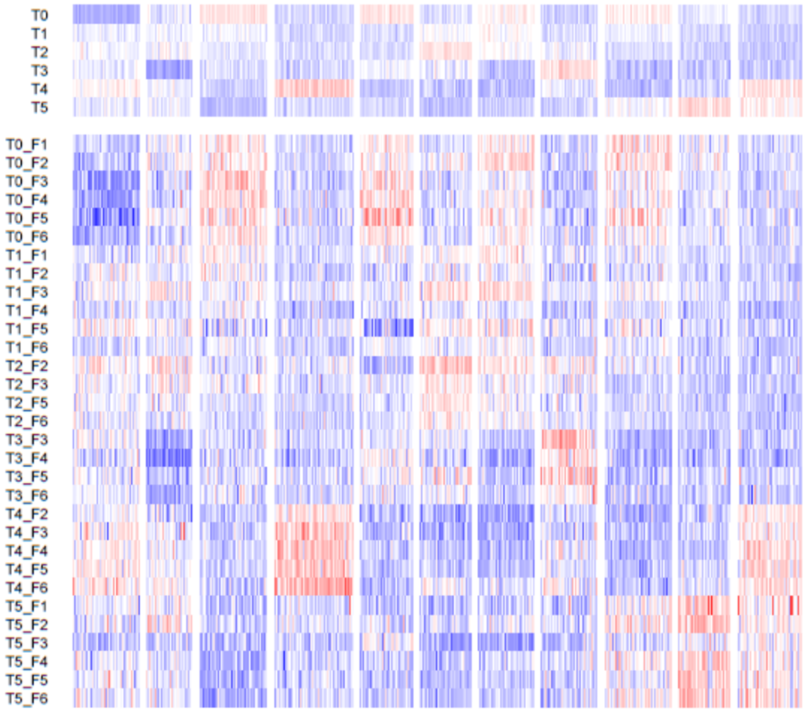


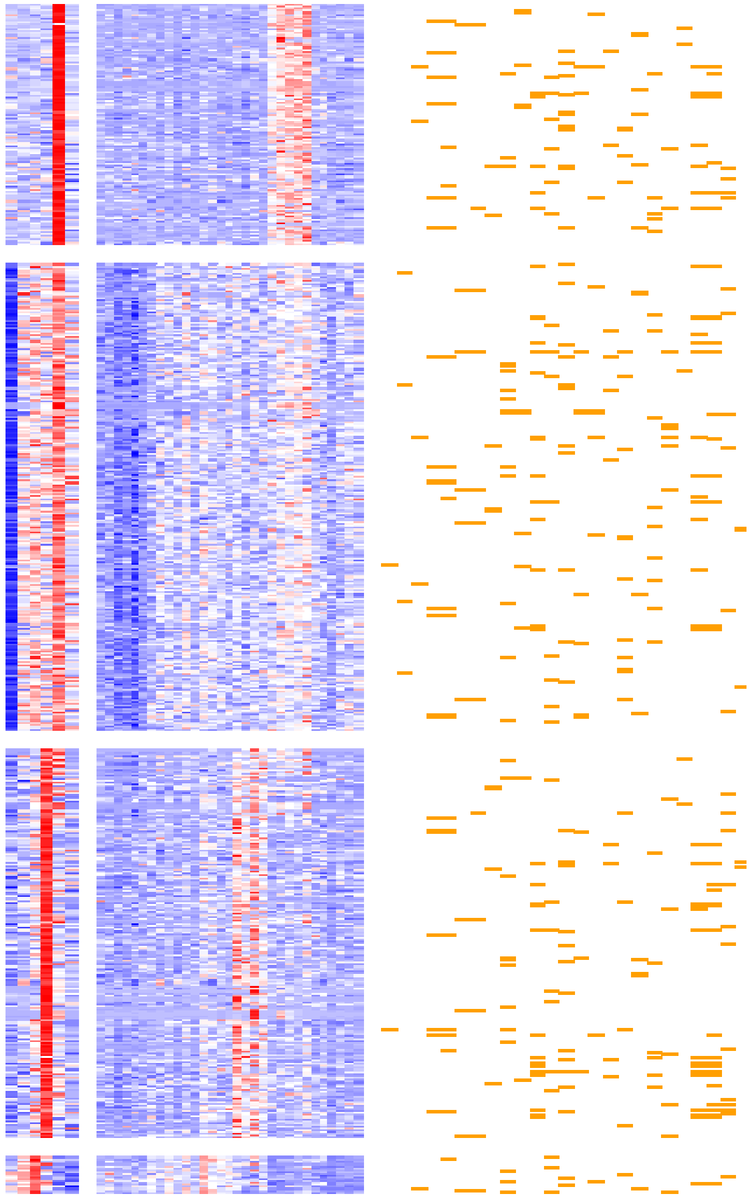
Anno_idx	ovary	testis	eyes	brain	gills	heart	muscle	liver	kidney	intestine	eggs	stage15	cell1	stage27	stage31	stage35
a63	122	13	126	380	76	245	52	559	192	11	75	8	30	23	15	2
a254	0	0	64	0	0	0	0	0	0	0	0	0	0	0	1	
a358	0	2	5	13	0	1	1	2	1	0	14	5	0	27	44	2
a80	897	1196	0	2	9	1	2	0	1	0	150	2	20	5	3	
a367	151	177	1194	2261	553	354	979	279	631	46	92	0	6	60	923	108
a131	10	8	77	68	41	69	25	137	93	6	0	3	0	77	37	10
a255	0	0	0	9	0	0	0	0	0	0	0	0	0	3	15	
a9	2	4	16	78	1	0	19	0	10	0	3	0	0	7	17	3
a199	5	5	25	12	31	30	27	21	195	3	0	8	1	57	64	5
a292	162	41	40	60	67	48	46	39	59	9	231	251897	201	194136	60277	2246

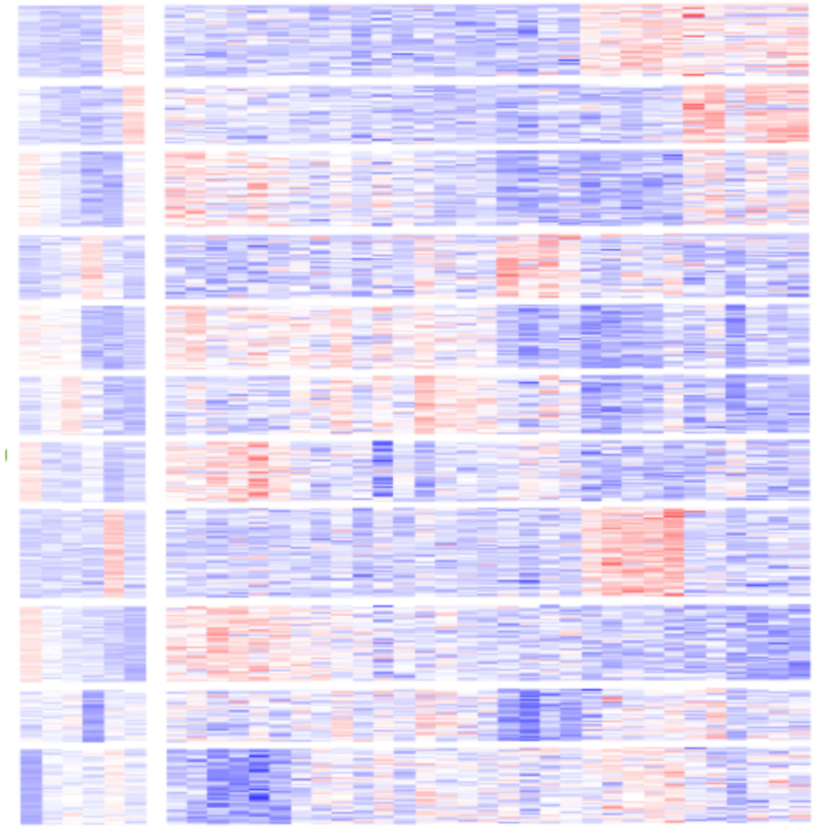
Applications

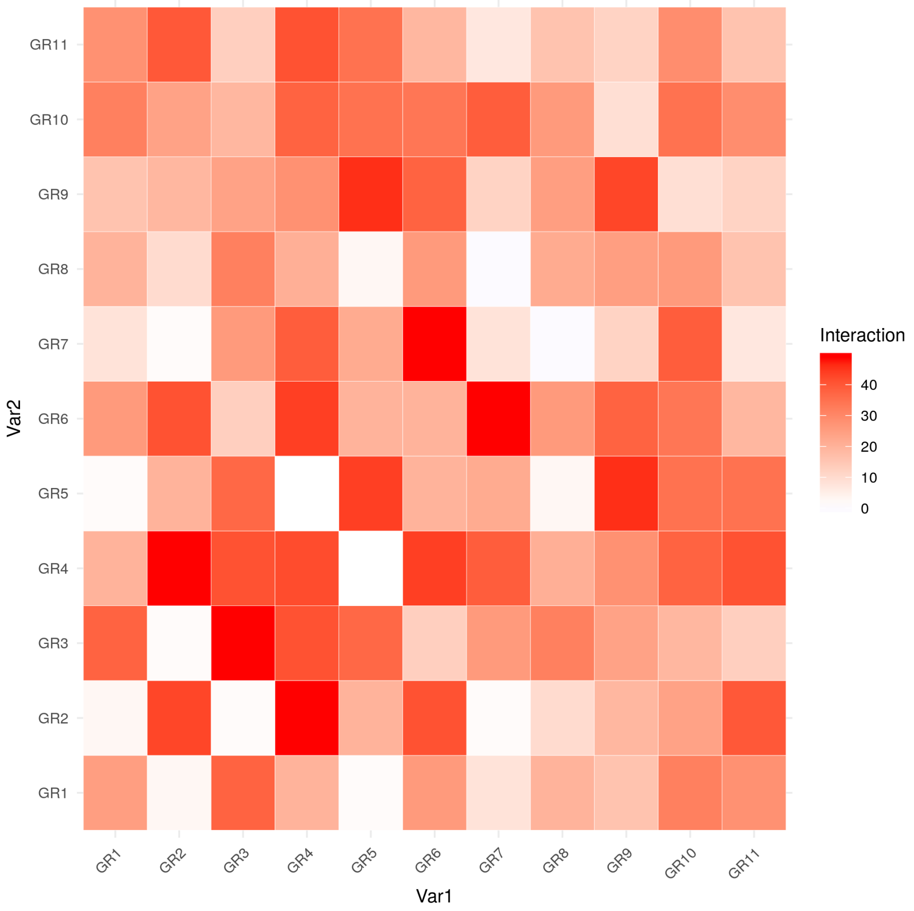
Interactions :

- Cis : feelnc classifier
- Trans : lncTar (Li et al. Brief Bioinform. 2015)



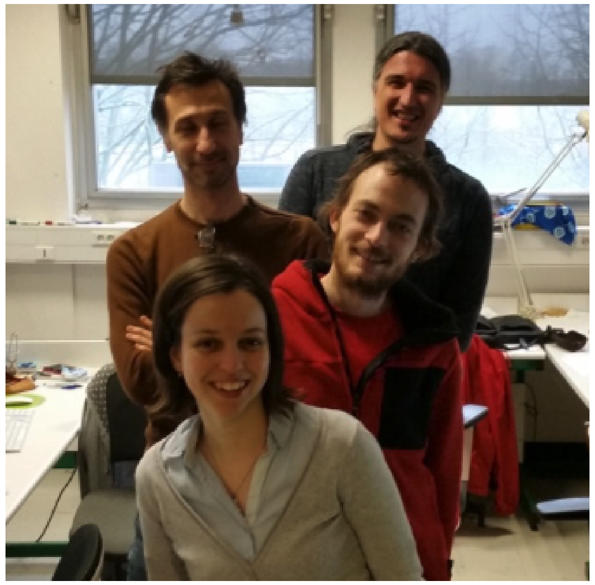






Remerciements

cnrs
dépasser les frontières



Guillaume Rizk

Inria

GENSCALE

Thomas Derrien
Valentin Wucher
Céline Le Béguec
Christophe Hitte

igdr

Fabrice Legeai
Susete Alves-Carvalho

Stephanie Robin
Cyril Monjeaud

IGEPP
Institut de
Génétique, Environnement
et Protection des Plantes

INRA
SCIENCE & IMPACT

GenOuest

BIPAA
BioInformatics Platform for Agroecosystem Arthropods