



TAGADA

A pipeline for flexible and scalable
quantification and annotation of transcripts

Cervin Guyomar, Cyril Kurylo, Sarah Djebali, Sylvain Foissac

Groupe Biopuces - 30/05/2024

The good old "Central dogma of molecular biology"



↑
Quantification of genes/transcripts using
- RNA-seq libraries
- a "reference annotation"

Reference annotations are critical

- Ensembl
- RefSeq (NCBI)
- UCSC

Comparing annotation sources

We developed **DRAGIBUS**, a set of metrics to evaluate the quality of a gtf/gff file

- Syntax / consistency check
- Number of genes/transcripts/exons
- Proportion of monoexonic transcripts
- Splice sites canonicity
- Occurrences of polyA signals in 3' UTRs

Comparing annotation sources

Number of genes/transcripts

	NCBI RefSeq	ENSEMBL	UCSC known genes	UCSC reference genes
galGal6 (chicken)	24K genes 62K transcripts	24K genes 39K transcripts	-	7K genes 7K transcripts
mm10 (mouse)	36K genes 107K transcripts	43K genes 104K transcripts	129K genes 142K transcripts	25K genes 44K transcripts
hg38 (human)	54K genes 191K transcripts	64K genes 208K transcripts	214K genes 248K transcripts	28K genes 78K transcripts

Comparing annotation sources

Proportion of monoexonic transcripts

	NCBI RefSeq	ENSEMBL	UCSC known genes	UCSC reference genes
galGal6 (chicken)	3.3%	7.4%	-	15.7%
oviAri3 (sheep)	3.5%	20.8%	-	14.3%

Comparing annotation sources

Transcripts with polyA signal support

	NCBI RefSeq	ENSEMBL	UCSC known genes	UCSC reference genes
mm10 (mouse)	52.2%	34.0%	32.4%	69.5%
hg38 (human)	45.4%	28.7%	29.7%	65.6%

Comparing annotation versions

Number of genes/transcripts

	ce6 (flatworm)	ce11 (flatworm)	bosTau6 (cow)	bosTau9 (cow)
ENSEMBL	28K genes 35K transcripts	47K genes 61K transcripts	25K genes 27K transcripts	28K genes 44K transcripts

Comparing species

Proportion of canonical splice sites

	hg38 (human)	mm10 (mouse)	bosTau9 (cow)	susScr11 (pig)	susScr3 (pig)
NCBI RefSeq	99.8%	99.8%	99.7%	99.4%	98.4%
ENSEMBL	99.8%	99.3%	99.5%	97.1%	89.5%

- Iterations of reference annotations are very labile
- Reference annotation pipelines are not explicitly documented and reproducible

Gene annotation in Ensembl

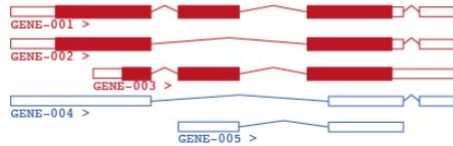
Gene annotation is the plotting of genes onto [genome assemblies](#), and indexing their genomic coordinates.

Gene annotation provided by Ensembl includes automatic annotation, ie genome-wide determination of transcripts. For selected species (ie human, mouse, zebrafish, rat), gene annotation may also include manual curation, ie reviewed determination of transcripts on a case-by-case basis. Furthermore, Ensembl imports annotation from FlyBase, WormBase and SGD.

Ensembl transcripts displayed on our website are products of the Ensembl automatic gene annotation system (a collection of gene annotation pipelines), termed the Ensembl annotation process. All Ensembl transcripts are based on experimental evidence and thus the automated pipeline relies on the mRNAs and protein sequences deposited into public databases from the scientific community. Manually-curated transcripts are produced by the HAVANA group.

An Ensembl gene (with a unique ENSG... ID) includes any spliced transcripts (ENST...) with overlapping coding sequence, with the exception of manually annotated readthrough genes which are annotated as a separate locus. Transcripts from the Ensembl annotation process, the Havana/Vega set and the Consensus Coding Sequence (CCDS project) set may all be clustered into the same gene. Transcripts that belong to the same gene ID may differ in transcription start and end sites, splice events and exons, and can give rise to very different proteins. Transcript clusters with no overlapping coding sequence are annotated as separate genes. Two transcripts may overlap in non-coding sequence (ie intronic sequence or UnTranslated Region (UTR)), and be classified under two separate genes. After the Ensembl gene and transcript sequences are defined, the gene and transcript names are assigned.

The image below shows a cartoon of a gene ("GENE") with five transcripts, some coding (red) and non-coding (blue).



The sequence of any gene or transcript shown in Ensembl is the sequence in the underlying genome assembly, where the sequence of any protein is the translated genomic sequence. This is to prevent any mismatch between the genes and the genome. For this reason, sequences of genes, transcripts and proteins in Ensembl may differ from other databases, who may use sequence from other individuals than were used to produce the genome.

Find out more about the different types of gene annotation used by Ensembl, and where we get our data from:

- [Automatic annotation of coding genes.](#)
- [Automatic annotation of non-coding genes.](#)
- [Annotation of immunoglobulin and T-cell receptor genes.](#)
- [Automatic annotation using RNA-seq data.](#)
- [Manual gene annotation by Havana.](#)
- [The Ensembl and Havana merge.](#)
- [MANE \(Matched annotation between NCBI and EBI\).](#)
- [CCDS.](#)
- [Gene annotation of low quality assemblies.](#)
- [Sources of data for gene annotation.](#)
- [Gene naming.](#)
- [Transcript tags.](#)
- [Gene and transcript types.](#)
- [External references.](#)

Annotation projects in our group

- FR-AgENCODE project (V2): 4 species, 4 animals, ~10 tissues
- GENE-SWitCH project: 2 species, 4 animals, 3 stages, ~7 tissues

FR-AgENCODE

Project overview Data & Results People Contact Related projects



FR-AgENCODE
A FAANG pilot project for the functional annotation of livestock genomes

Download the new chicken annotation

3 molecular assays

Chromatin accessible gene regulatory elements (ATAC-seq)

Long range interactions (HiC/4C)

RNA polymerase

Expression RNA-seq/RNA-seq

14 research labs & facilities



4 livestock species

Bos taurus Holstein

Capra hircus Alpine

Gallus gallus White Leghorn

Sus scrofa Large White

www.fragencode.org



Functional annotation of 7 tissues during development

- 3 developmental stages: Early organogenesis



Late organogenesis



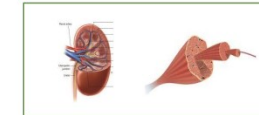
Newborns



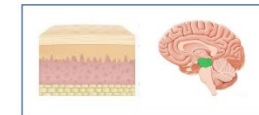
- 7 Tissues: liver, skeletal muscle, small intestine, cerebellum, dorsal epidermis, lung and kidney



Endoderm



Mesoderm



Ectoderm

www.gene-switch.eu

The TAGADA pipeline

TAGADA : all the goodness of **nextflow**



Portable

Docker + Singularity



Scalable

Slurm, Kubernetes....



Modular

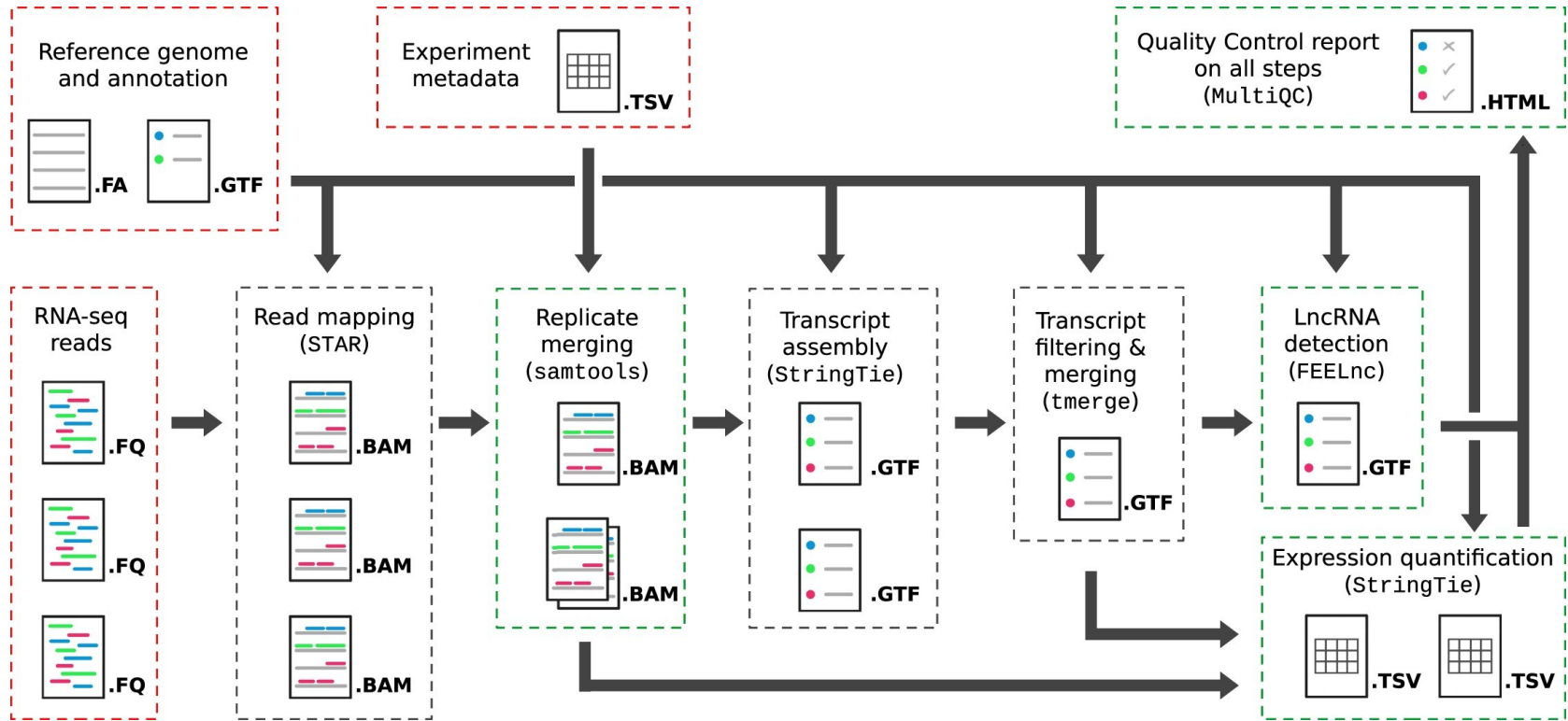
DSL2 processes



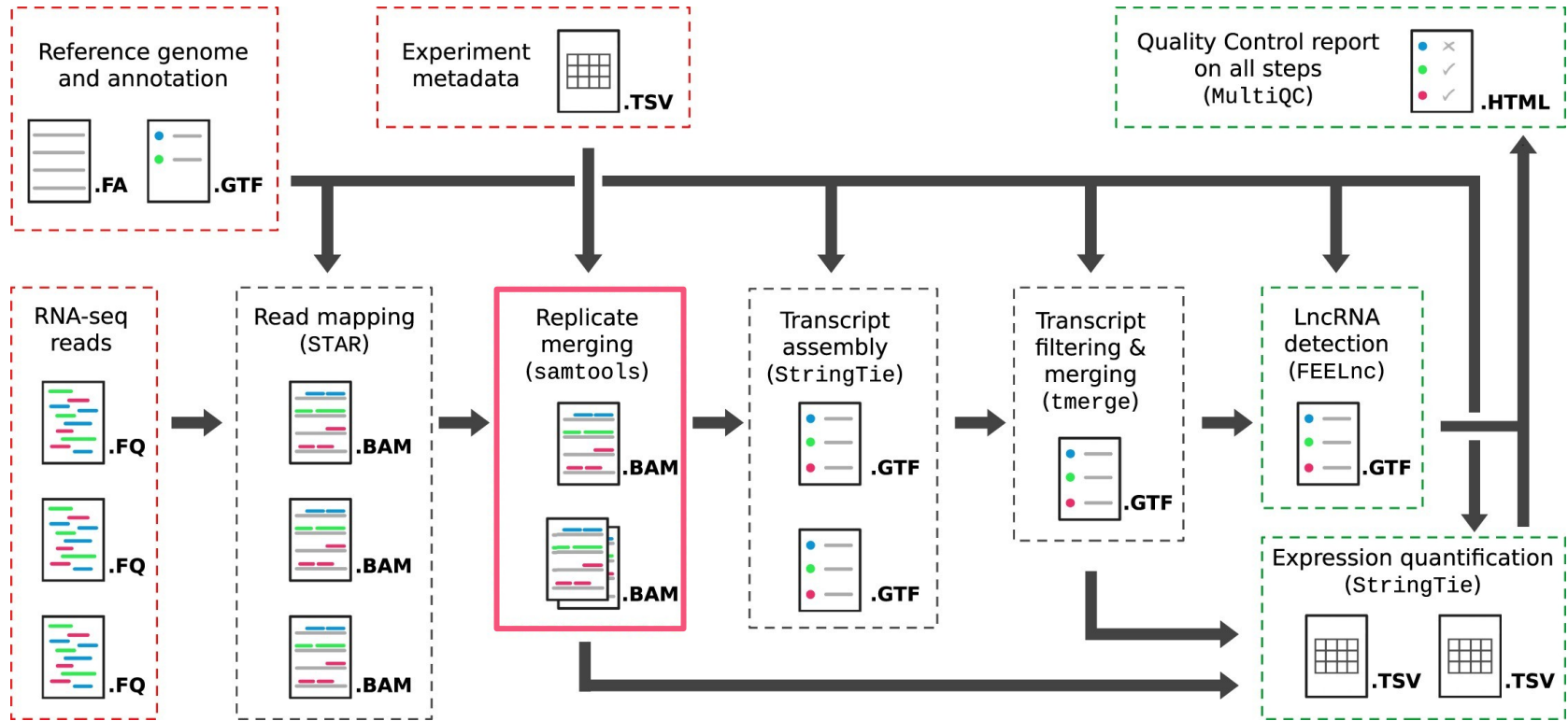
Reproducible

```
nextflow run FAANG/analysis-TAGADA \  
-profile test,docker \  
-revision 2.1.3 --output directory
```

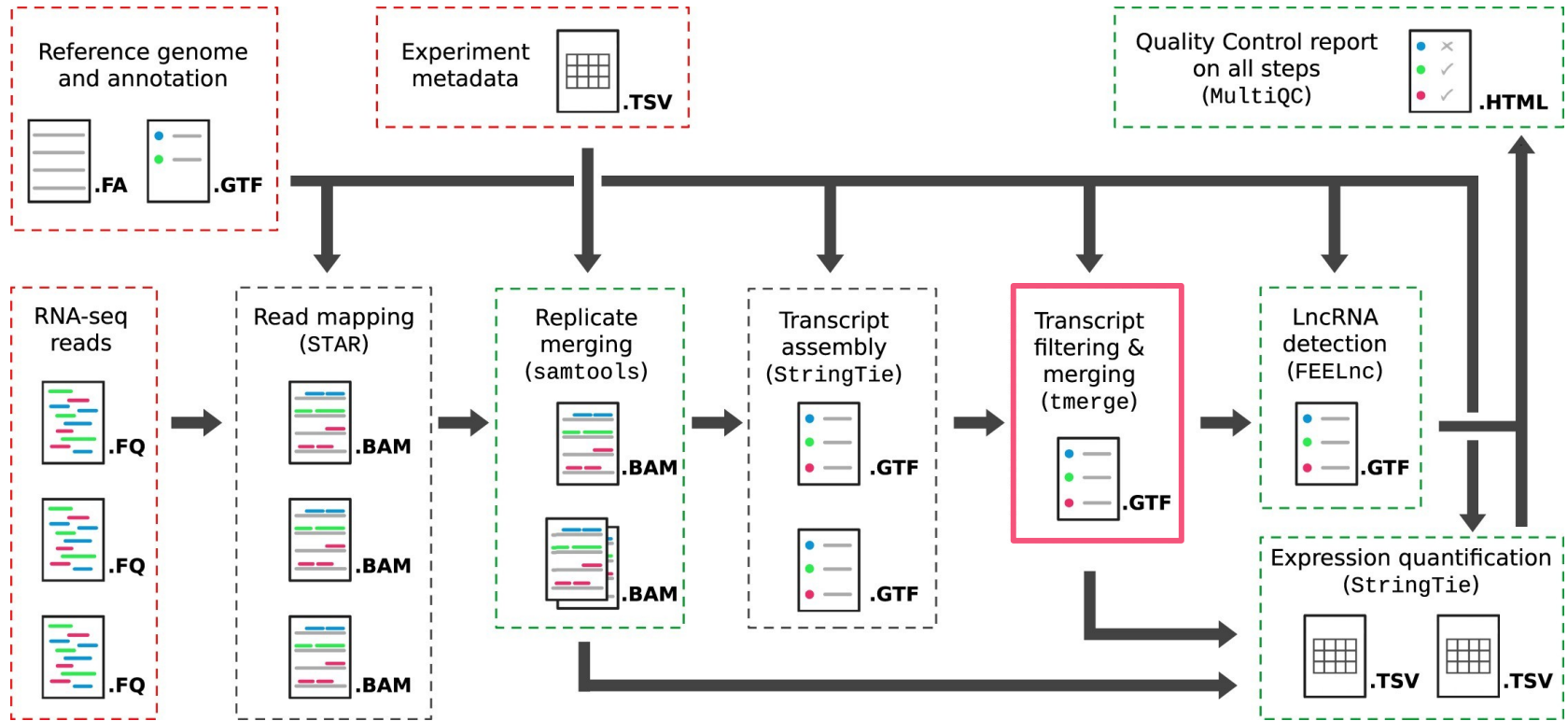
TAGADA : the pipeline



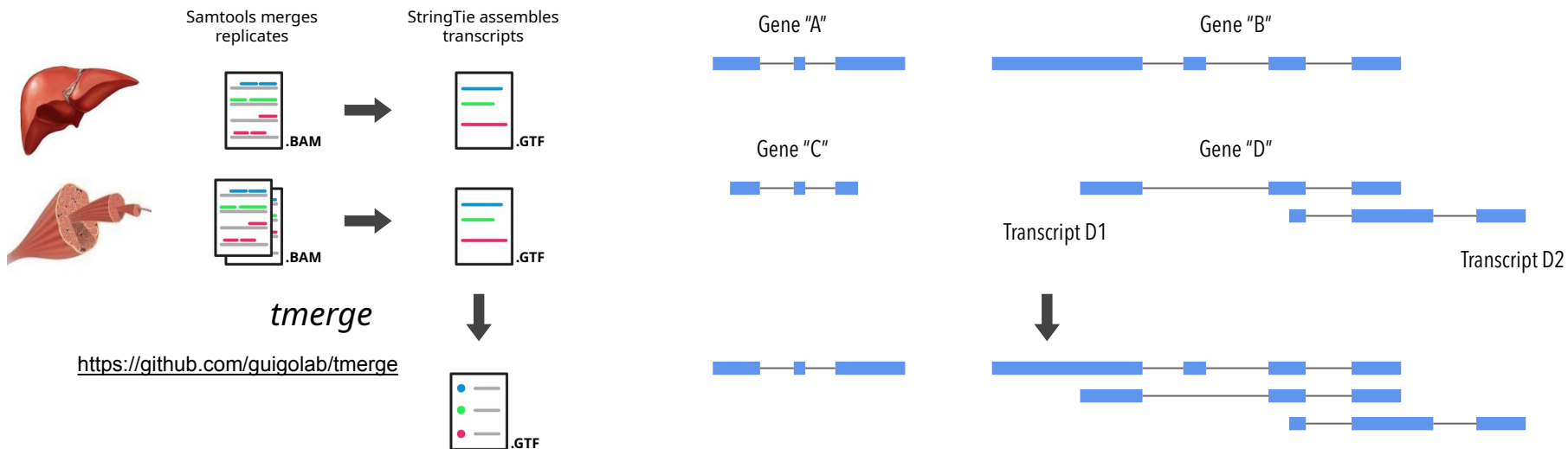
Flexible replicate management



Merging annotations



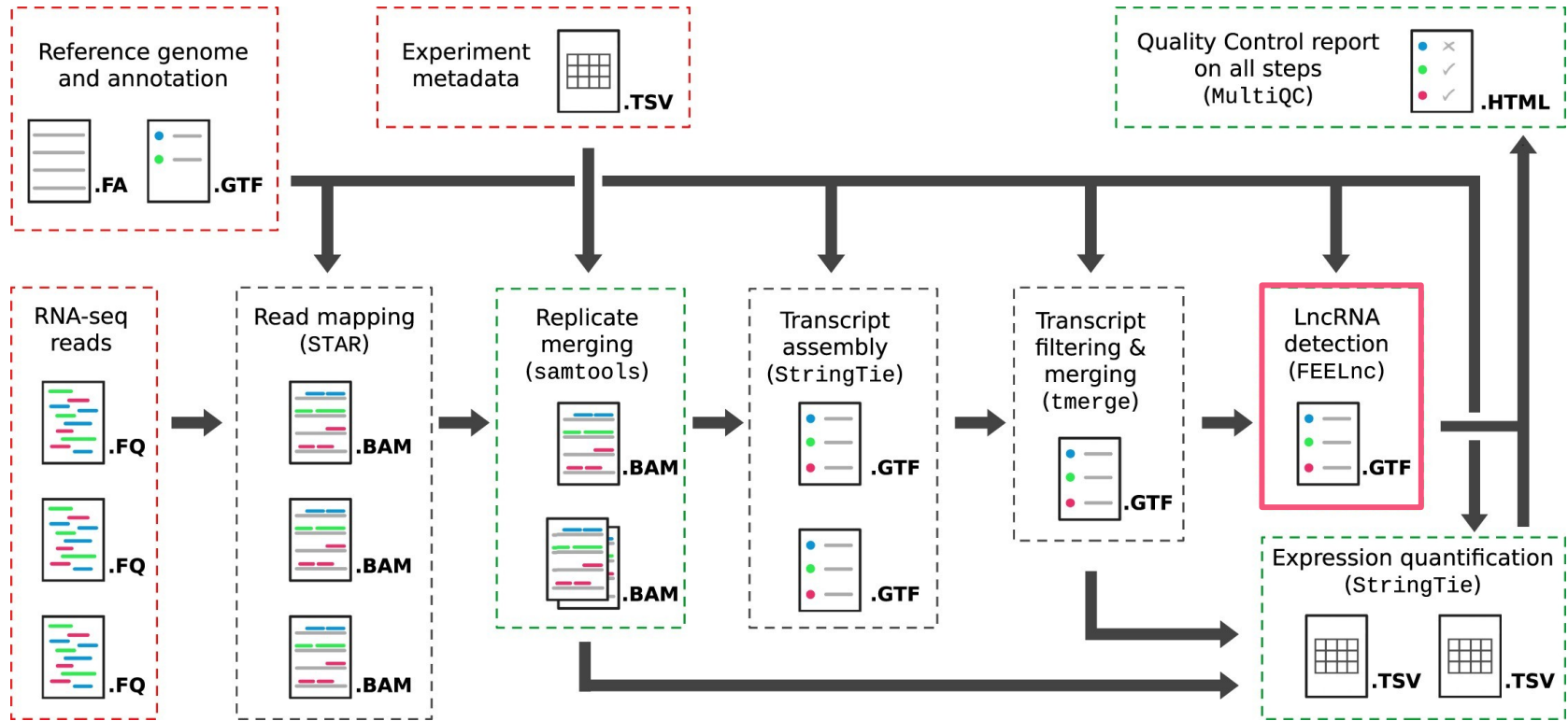
Merging annotations





Additional features :

- expression-based filtering
- consistency-based filtering
- reference annotation inclusion

LncRNA annotation



TAGADA features vs nf-core

	Generic metadata	Aligner	Transcripts assembler	Transcripts assembly	Transcripts coalescer	Novel annotation	Non-coding transcripts detection	Quantifier	Quantification
nf-core/ rnaseq 	NO	STAR or HISAT2 or Salmon	StringTie	Transcripts assembly in each replicate group	NO	Individual annotation for each replicate group	NO	Reference annotation	
								RSEM or Salmon	Individual matrix for each replicate group with TPM and read counts
								Novel annotation	
								StringTie	Individual matrix for each replicate group with TPM but no read counts
FAANG/ analysis-TAGADA 	YES	STAR	StringTie	Transcripts assembly in any group level	StringTie or Tmerge	One annotation	FEELnc	Reference annotation	
								StringTie	One matrix with TPM and read counts
								Novel annotation	
								StringTie	One matrix with TPM and read counts

GENESWITCH TAGADA results

- Gene/transcript quantification
- Novel gene/transcript annotation

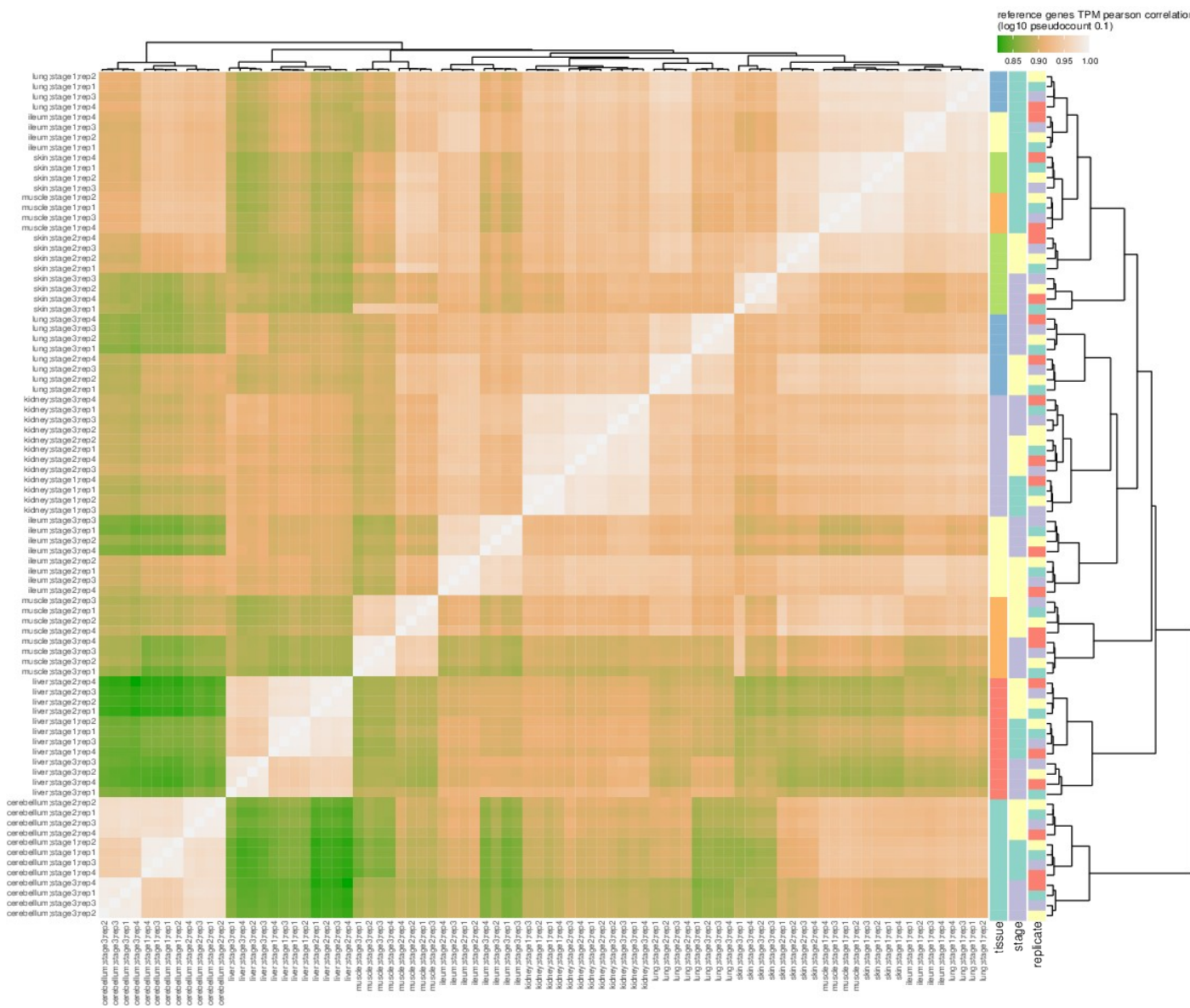
GENESWITCH data

🍓 84 experiments (7 tissues x 3 developmental stages x 4 animals) per species:

- PolyA+ RNAs
- Directional, PE 150 sequencing
- +100-150 million PE reads / experiment

🍓 Reference genomes and gene annotations:

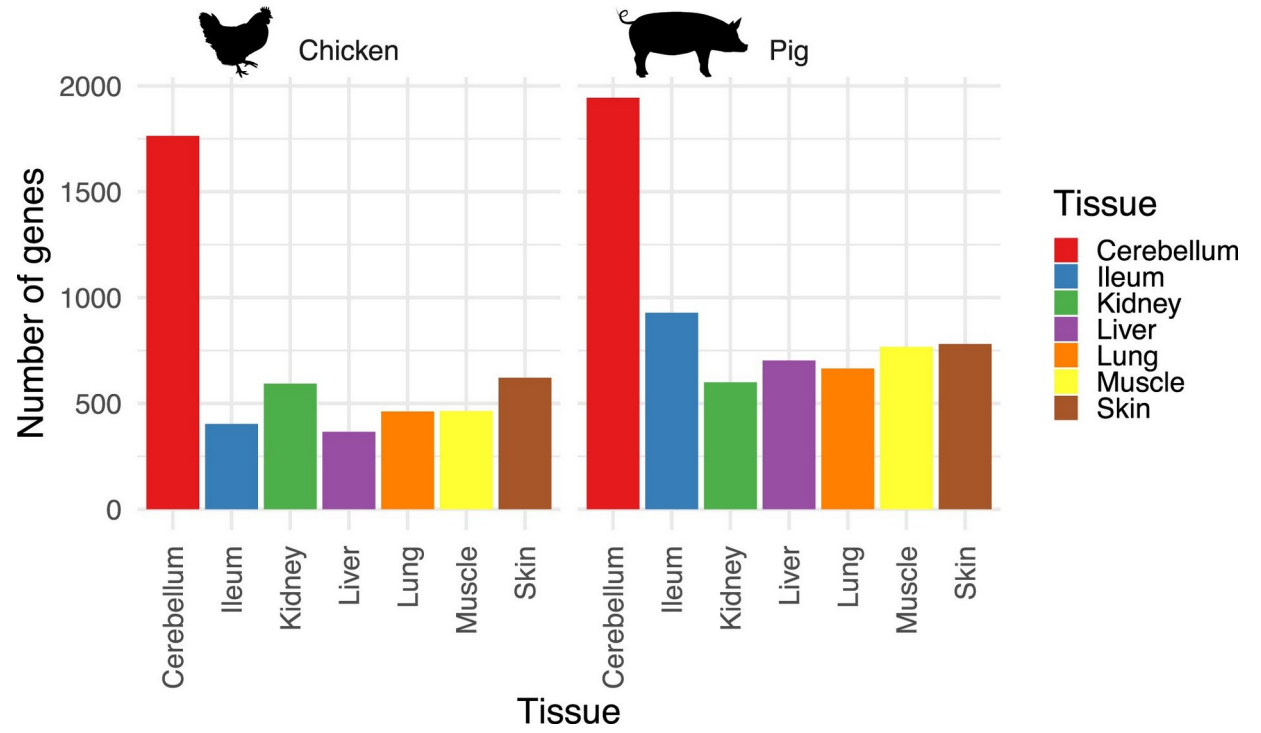
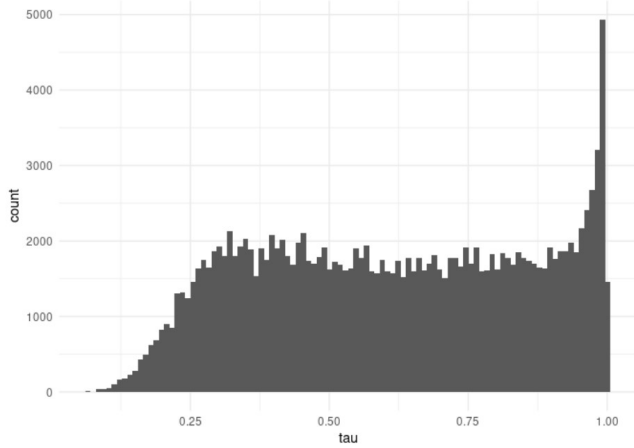
- Chicken:
 - Galgal6, ensembl v102
- Pig:
 - Sscrofa11.1, ensembl v102



Clustering of reference genes quantification (pig)

Tissue specific genes using the Tau index


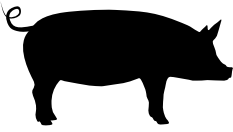
$$\tau = \frac{\sum_{i=1}^n (1 - \hat{x}_i)}{n-1}; \quad \hat{x}_i = \frac{x_i}{\max_{1 \leq i \leq n} (x_i)}$$



Functional annotation of tissue specific genes



Novel genome annotation

	# reference genes	# TAGADA genes	# reference transcripts	# TAGADA transcripts
	24,356	34,712	39,288	138,272
	31,908	51,171	63,041	197,396

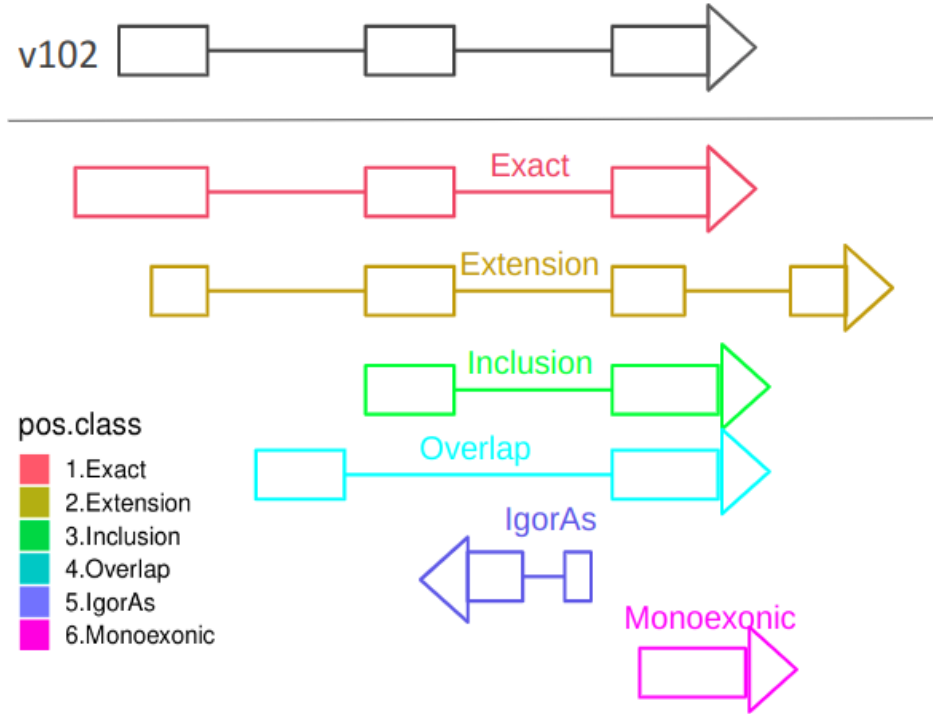
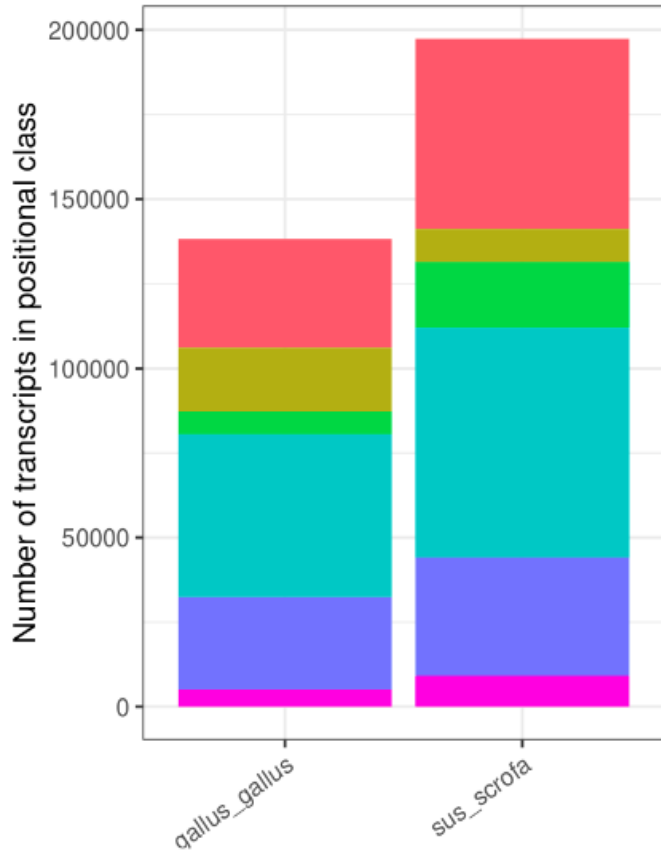


X 1.4-1.6 genes

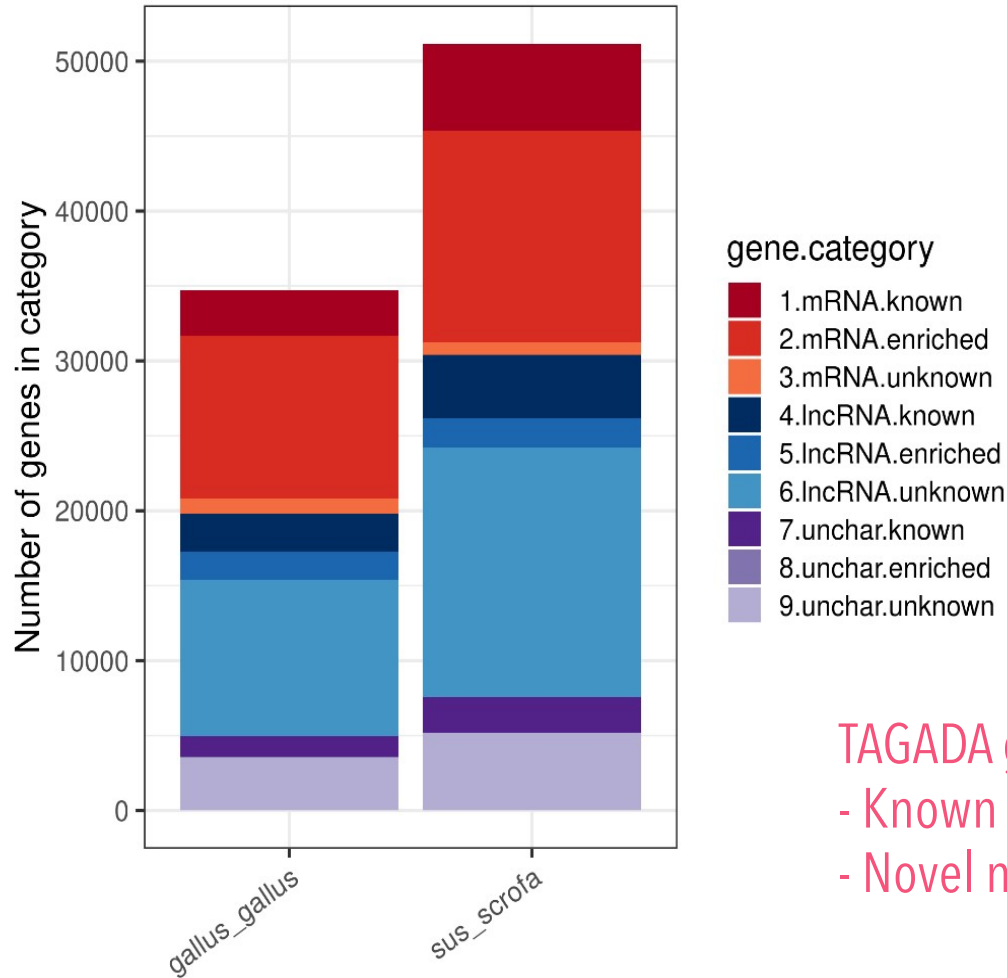


x 3.1-3.5 transcripts

What are the new transcripts - structure



What are the new transcripts – coding status



TAGADA genes are :

- Known coding genes with novel transcripts
- Novel non coding genes

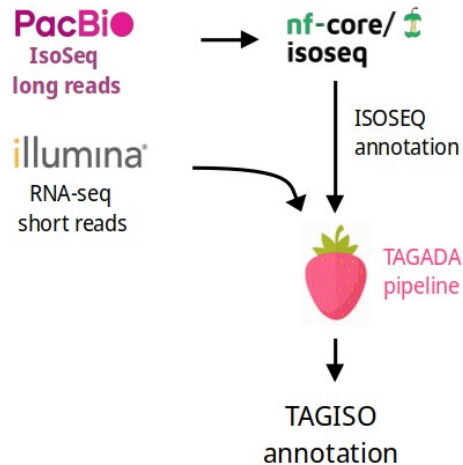
DRAGIBUS evaluation of TAGADA transcripts

Species	Gene annotation	canonical splice sites	TSS with ATAC-seq support	internal exons > 500bp	TTS with polyA site
Chicken	TAGADA	99.5%	73.9%	6.1%	34.7%
Pig	ensv102	98.3%	67.0%	2.4%	43.0%
	TAGADA	98.9%	63.6%	8.3%	32.5%
	ensv102	97.1%	64.0%	5.1%	42.8%

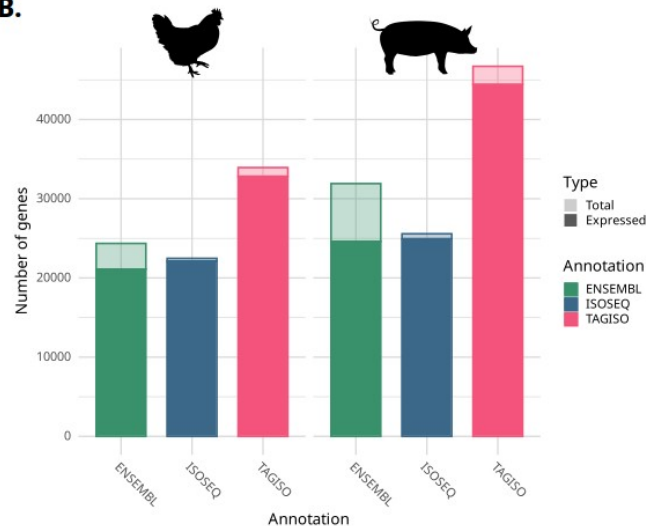


Bonus feature : Combine long-read annotation and RNA-seq

A.



B.



Species	Annotation	TSS ATAC-seq support
chicken	ensv102	67.0%
	Tagada on ensv102	73.9%
	Tagada on long reads	78.6%
pig	ensv102	64.0%
	Tagada on ensv102	63.6%
	Tagada on long reads	67.8%

Take-Home messages

- 🍓 TAGADA is fast and easy to use pipeline to annotate and quantify transcripts
- 🍓 TAGADA was applied on the GENE-SWItCH RNA-seq data to produce a transcriptome profiling and an extended annotation with good properties
- 🍓 TAGADA can integrate RNA-seq reads with an long-read annotation, improving the resulting annotation



<https://github.com/FAANG/analysis-TAGADA>



NAR-GAB article

Thank you for your attention



Acknowledgements

INRAE/INSERM Toulouse

- Sarah Djebali
- Sylvain Foissac
- Cyril Kurylo

Roslin Institute

- Sebastien Guizard

GENE-SwitCH

- Hervé Acloque
- Alan Archibald
- Elisabetta Giuffra

