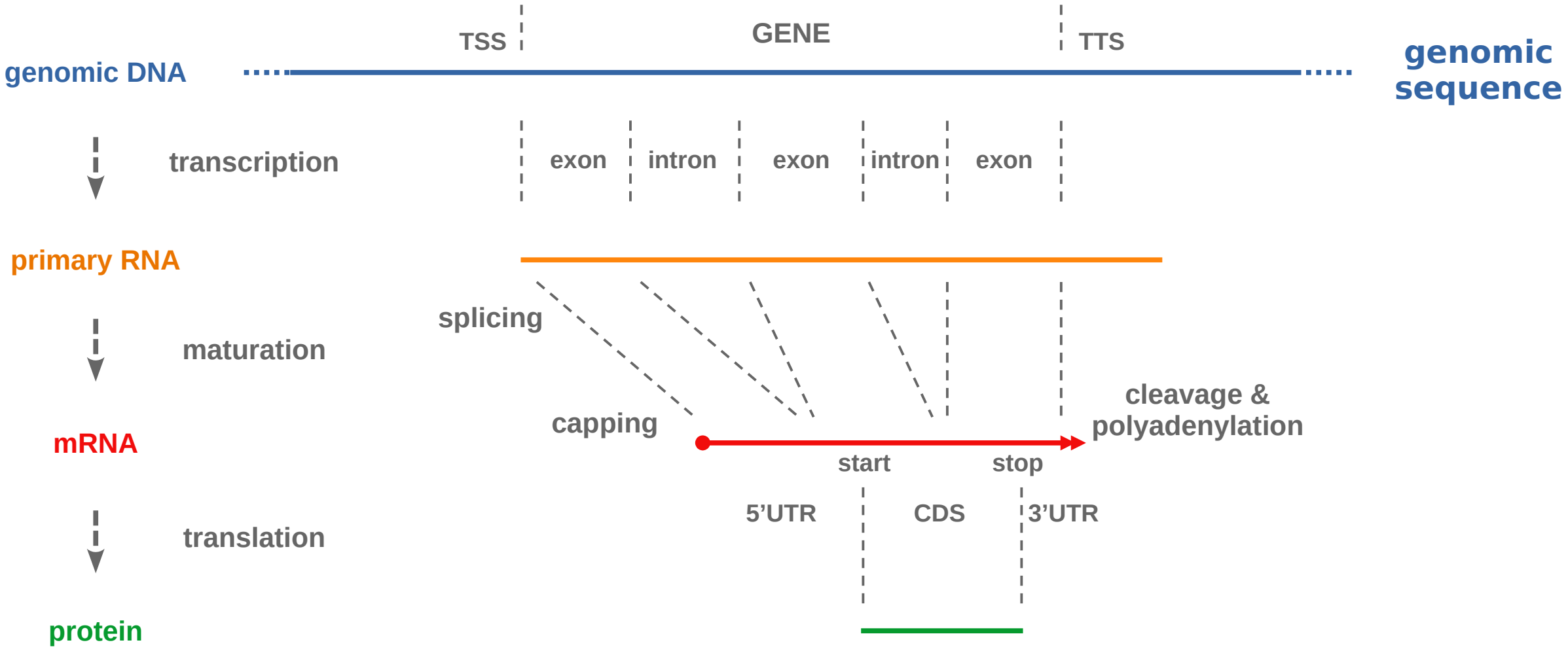
Annotating genes and transcripts in animal genomes

From RNA-seq data to gene and transcript annotations

Sylvain Foissac, INRAE, June 2025

Context & preliminary definitions

Context & preliminary definitions



Eukaryotic gene expression: historical & oversimplified model

Context & preliminary definitions

- “**Genome**”: reference genomic DNA sequence of the species. Here: vertebrates (livestock)
  Questions: version? (pan)genome? chromosomes? haplotypes? (un)masked?

Context & preliminary definitions

- “**Genome**”: reference genomic DNA sequence of the species. Here: vertebrates (livestock)
  Questions: version? (pan)genome? chromosomes? haplotypes? (un)masked?
  Genomic funfacts:
  - about half of your genome is repeats
  - all your cells do not have the same genomic sequence (somatic mutation)
  - ~70% of your cardiac cells and ~30% of your hepatocytes contain polyploid nuclei (≥4n)

(Donne et al, Nat Rev gast & hep, 2020 ; Kapranov et al, BMC Biology, 2010; Cappannini et al, NAR, 2023 ; Wikipedia.org)

Context & preliminary definitions

- “**Genome**”: reference genomic DNA sequence of the species. Here: vertebrates (livestock)
  Questions: version? (pan)genome? chromosomes? haplotypes? (un)masked?
  Genomic funfacts:
  - about half of your genome is repeats
  - all your cells do not have the same genomic sequence (somatic mutation)
  - ~70% of your cardiac cells and ~30% of your hepatocytes contain polyploid nuclei (≥4n)
- **Transcript**: RNA molecule generated by transcribing a genomic fragment.
  Questions: coding or non-coding? long or small? capped/spliced/polyAed?

(Donne et al, Nat Rev gast & hep, 2020 ; Kapranov et al, BMC Biology, 2010; Cappannini et al, NAR, 2023 ; Wikipedia.org)

Context & preliminary definitions

- "**Genome**": reference genomic DNA sequence of the species. Here: vertebrates (livestock)
  Questions: version? (pan)genome? chromosomes? haplotypes? (un)masked?
  Genomic funfacts:
  - about half of your genome is repeats
  - all your cells do not have the same genomic sequence (somatic mutation)
  - ~70% of your cardiac cells and ~30% of your hepatocytes contain polyploid nuclei (≥4n)

- **Transcript**: RNA molecule generated by transcribing a genomic fragment.
  Questions: coding or non-coding? long or small? capped/spliced/polyAed?
  Transcriptome funfacts:
  - ~95% of the RNAs in your cells are small and non-coding (rRNAs)
  - most of the non-ribosomal RNA content in your cell is non-coding (ncRNA)
  - ~80% of the transcribed DNA used to produce a coding RNA is non-coding (intron+UTR)
  - there are ~170 documented RNA epitranscriptomic modifications (m^6A, m^5C, Ψ, etc)
  - transcripts do exist.

(Donne et al, Nat Rev gast & hep, 2020 ; Kapranov et al, BMC Biology, 2010; Cappannini et al, NAR, 2023 ; Wikipedia.org)

Context & preliminary definitions

- “**Gene**”: no (consensus) definition

Stadler (1954): “The Gene” - Demerec (1955): “What is a Gene ? - Twenty Years Later” - Portin (1993): “The Concept of the Gene : Short History and Present Status” - Snyder & Gerstein (2003): “Defining Genes in the Genomics Era” - Pearson (2006): “What is a gene ?” - Pennisi (2007): “DNA Study Forces Rethink of What It Means to Be a Gene” - Fox Keller & Harel (2007): “Beyond the Gene” - Gerstein et al (2007): “What is a gene, post-ENCODE ? History and updated definition” - Pesole (2008): “What is a gene ? An updated operational definition” - Stadler (2009): “Defining genes : a computational framework” - Portin & Wilkins (2017): “The Evolving Definition of the Term ‘Gene’”

Context & preliminary definitions

- "**Gene**": no (consensus) definition
  - Mendeleian definition: basic unit of heredity (heritable trait)
  - Molecular definition: DNA sequence that codes for a (functional?) molecular product, RNA and/or protein
  - "Minimal" definition: genomic interval with reported evidence of transcriptional activity
  - "Bioinformatics" practical definition: genomic interval with overlapping transcripts (connected component of annotated transcripts with ≥1 bp sense exonic overlap)

Stadler (1954): "The Gene" - Demerec (1955): "What is a Gene ? - Twenty Years Later" - Portin (1993): "The Concept of the Gene : Short History and Present Status" - Snyder & Gerstein (2003): "Defining Genes in the Genomics Era" - Pearson (2006): "What is a gene ?" - Pennisi (2007): "DNA Study Forces Rethink of What It Means to Be a Gene" - Fox Keller & Harel (2007): "Beyond the Gene" - Gerstein et al (2007): "What is a gene, post-ENCODE ? History and updated definition" - Pesole (2008): "What is a gene ? An updated operational definition" - Stadler (2009): "Defining genes : a computational framework" - Portin & Wilkins (2017): "The Evolving Definition of the Term 'Gene'"

Context & preliminary definitions

- "**Gene**": no (consensus) definition
  - Mendeleian definition: basic unit of heredity (heritable trait)
  - Molecular definition: DNA sequence that codes for a (functional?) molecular product, RNA and/or protein
  - "Minimal" definition: genomic interval with reported evidence of transcriptional activity
  - "Bioinformatics" practical definition: genomic interval with overlapping transcripts (connected component of annotated transcripts with ≥1 bp sense exonic overlap)

Gene funfacts:

- the majority of the human genes are non-coding
- various transcripts can be produced from any genomic position (pervasive transcription, alternative events, back/trans/recursive-splicing, cutting/recapping, RNA copying...)
- genes cannot be defined, counted, delimited, completed

Stadler (1954): "The Gene" - Demerec (1955): "What is a Gene ? - Twenty Years Later" - Portin (1993): "The Concept of the Gene : Short History and Present Status" - Snyder & Gerstein (2003): "Defining Genes in the Genomics Era" - Pearson (2006): "What is a gene ?" - Pennisi (2007): "DNA Study Forces Rethink of What It Means to Be a Gene" - Fox Keller & Harel (2007): "Beyond the Gene" - Gerstein et al (2007): "What is a gene, post-ENCODE ? History and updated definition" - Pesole (2008): "What is a gene ? An updated operational definition" - Stadler (2009): "Defining genes : a computational framework" - Portin & Wilkins (2017): "The Evolving Definition of the Term 'Gene'" – Kurylo et al, 2023 – "The complete sequence of a human genome" Nurk et al, Science, 2022

Context & preliminary definitions

- **Genomic annotation**: document type and position of “functional” elements in a genome

  Questions: structural or functional annotation? Genes, transcripts, regulators (enhancers & silencers), repeats, others? What kind of information to use: DNA sequence, transcriptome/proteome, homology, previous annotation?

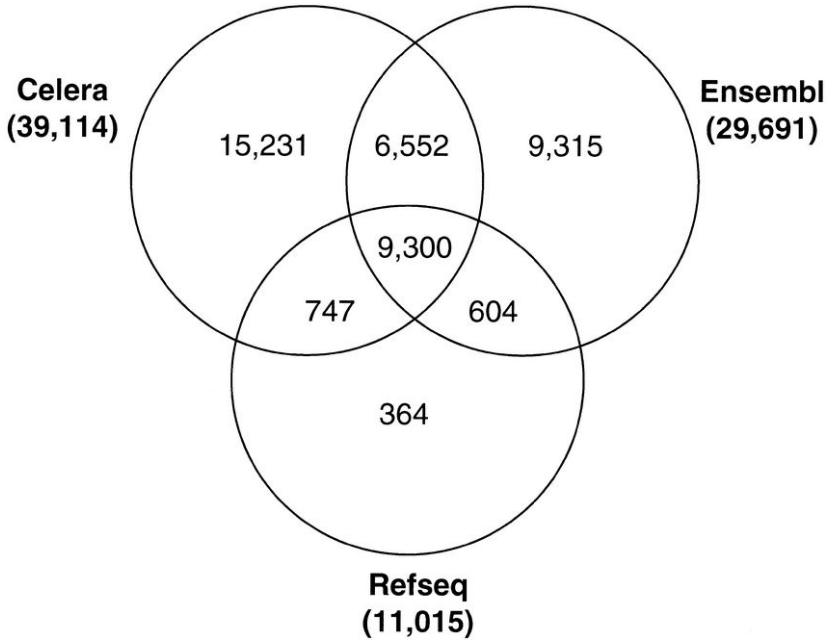
Context & preliminary definitions

- **Genomic annotation**: document type and position of “functional” elements in a genome

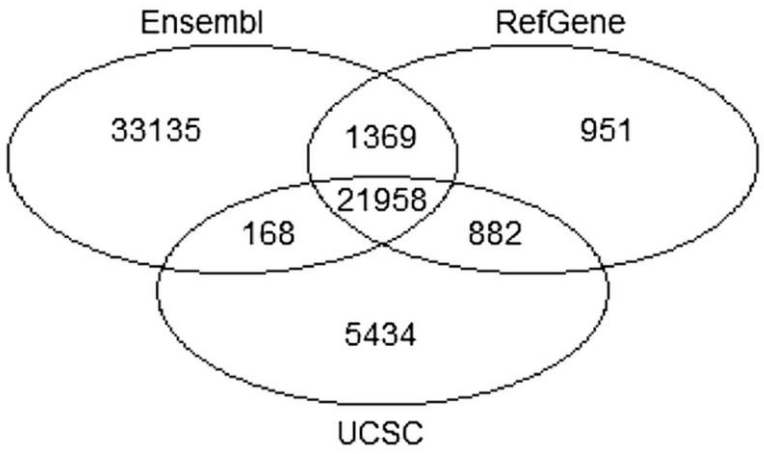
  Questions: structural or functional annotation? Genes, transcripts, regulators (enhancers & silencers), repeats, others? What kind of information to use: DNA sequence, transcriptome/proteome, homology, previous annotation?

  “Reference” annotations funfacts

  - Many sources: NCBI RefSeq, EMBL/EBI ENSEMBL, UCSC, GENCODE, FlyBase, TAIR, etc
  - inconsistency prevails



(Hogenesch et al. Cell, 2001)



(Zhao and Zhang BMC Genomics, 2015)

Context & preliminary definitions

- **Genomic annotation**: document type and position of “functional” elements in a genome

  Questions: structural or functional annotation? Genes, transcripts, regulators (enhancers & silencers), repeats, others? What kind of information to use: DNA sequence, transcriptome/proteome, homology, previous annotation?

  “Reference” annotations funfacts

  - Many sources: NCBI RefSeq, EMBL/EBI ENSEMBL, UCSC, GENCODE, FlyBase, TAIR, etc
  - inconsistency prevails

	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

Context & preliminary definitions

- **Genomic annotation**: document type and position of “functional” elements in a genome

  Questions: structural or functional annotation? Genes, transcripts, regulators (enhancers & silencers), repeats, others? What kind of information to use: DNA sequence, transcriptome/proteome, homology, previous annotation?

  “Reference” annotations funfacts

  - Many sources: NCBI RefSeq, EMBL/EBI ENSEMBL, UCSC, GENCODE, FlyBase, TAIR, etc
  - inconsistency prevails

	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

Context & preliminary definitions

- **Genomic annotation**: document type and position of “functional” elements in a genome

  Questions: structural or functional annotation? Genes, transcripts, regulators (enhancers & silencers), repeats, others? What kind of information to use: DNA sequence, transcriptome/proteome, homology, previous annotation?

  “Reference” annotations funfacts

  - Many sources: NCBI RefSeq, EMBL/EBI ENSEMBL, UCSC, GENCODE, FlyBase, TAIR, etc
  - inconsistency prevails

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%

Context & preliminary definitions

- **Genomic annotation**: document type and position of “functional” elements in a genome

  Questions: structural or functional annotation? Genes, transcripts, regulators (enhancers & silencers), repeats, others? What kind of information to use: DNA sequence, transcriptome/proteome, homology, previous annotation?

  “Reference” annotations funfacts

  - Many sources: NCBI RefSeq, EMBL/EBI ENSEMBL, UCSC, GENCODE, FlyBase, TAIR, etc
  - inconsistency prevails

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%

Context & preliminary definitions

- **Genomic annotation**: document type and position of “functional” elements in a genome

  Questions: structural or functional annotation? Genes, transcripts, regulators (enhancers & silencers), repeats, others? What kind of information to use: DNA sequence, transcriptome/proteome, homology, previous annotation?

  “Reference” annotations funfacts

  - Many sources: NCBI RefSeq, EMBL/EBI ENSEMBL, UCSC, GENCODE, FlyBase, TAIR, etc
  - inconsistency prevails

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%**

Context & preliminary definitions

- **Genomic annotation**: document type and position of “functional” elements in a genome

  Questions: structural or functional annotation? Genes, transcripts, regulators (enhancers & silencers), repeats, others? What kind of information to use: DNA sequence, transcriptome/proteome, homology, previous annotation?

  “Reference” annotations funfacts

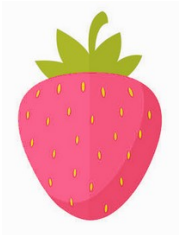
  - Many sources: NCBI RefSeq, EMBL/EBI ENSEMBL, UCSC, GENCODE, FlyBase, TAIR, etc
  - inconsistency prevails
  - generated by expert groups using “black box” annotation pipelines
    - not automatic
    - non reproducible
    - closed source
    - input unclear

need to improve annotation pipelines

Objective

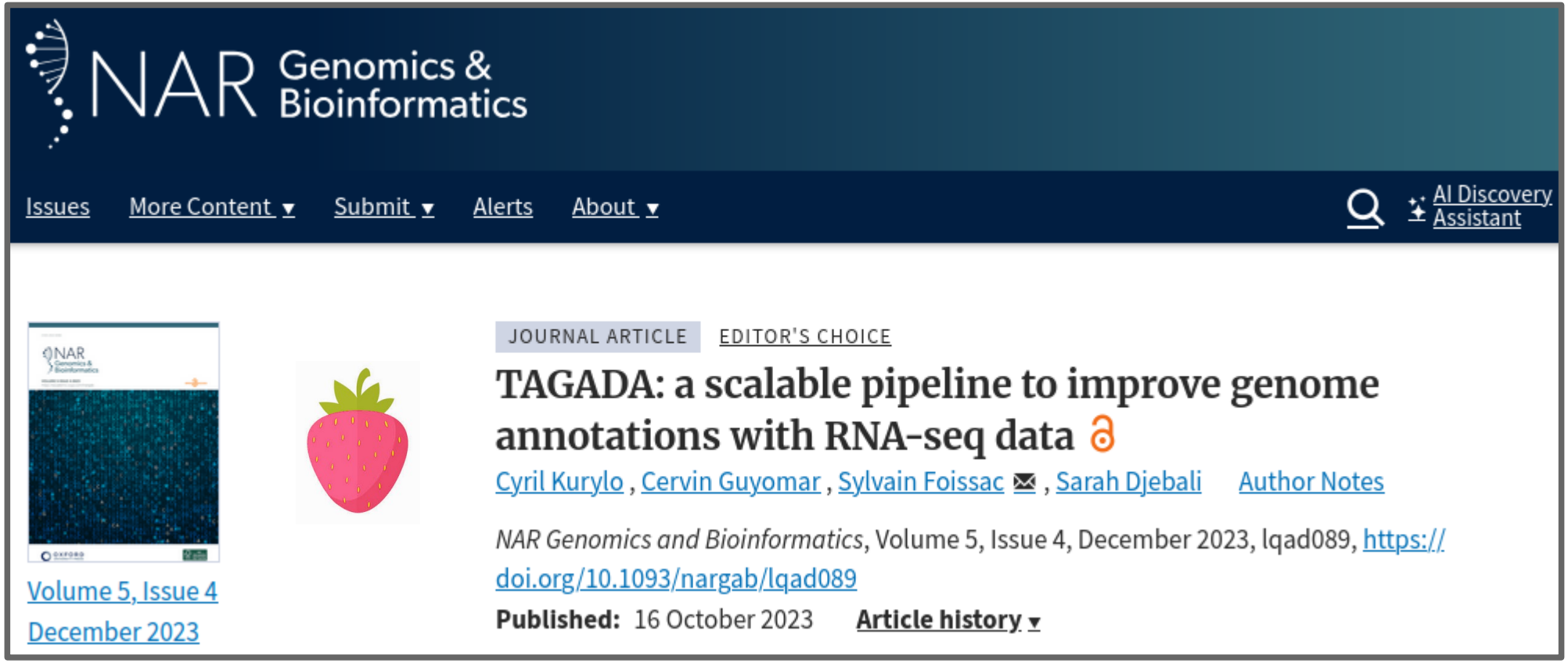
A genome annotation pipeline able to

- use RNA-seq data to improve an existing annotation (not just quantify expression)
- generate a new consensus annotation by merging results from various samples
- scale up with many samples and replicates (no “inflation”)
- quantify gene/transcript expression (read counts) in both reference and new annotations
- be portable, reproducible, open-source, user-friendly, FAIR
- predict transcript coding status (mRNA/lncRNA)
- start from fastq or bam
- propose many QC, filtering options



The TAGADA pipeline

Transcript And Gene Annotation, Deconvolution, Analysis



NAR Genomics & Bioinformatics

Issues More Content Submit Alerts About AI Discovery Assistant

Volume 5, Issue 4
December 2023

JOURNAL ARTICLE EDITOR'S CHOICE

TAGADA: a scalable pipeline to improve genome annotations with RNA-seq data

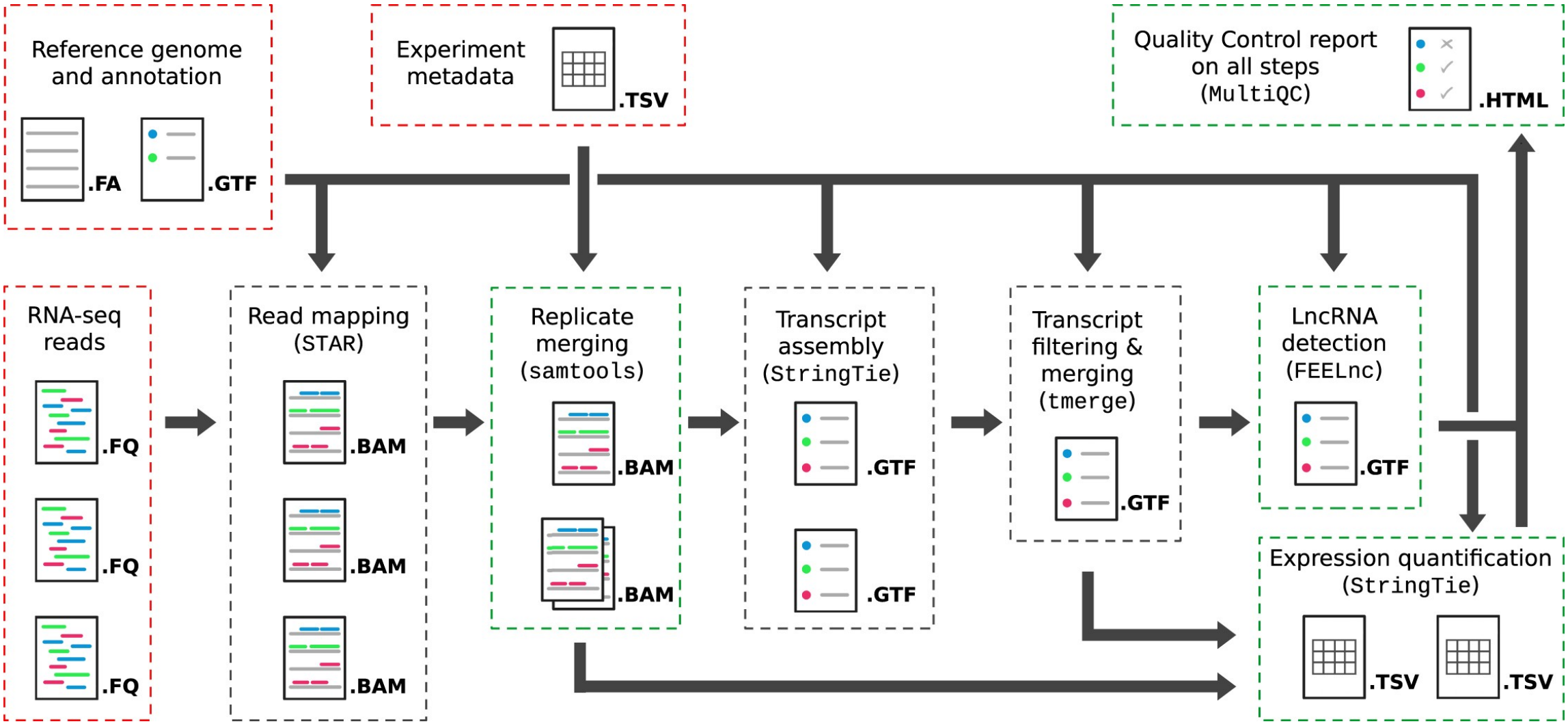
Cyril Kurylo , Cervin Guyomar , Sylvain Foissac , Sarah Djebali Author Notes

NAR Genomics and Bioinformatics, Volume 5, Issue 4, December 2023, lqad089, https://doi.org/10.1093/nargab/lqad089

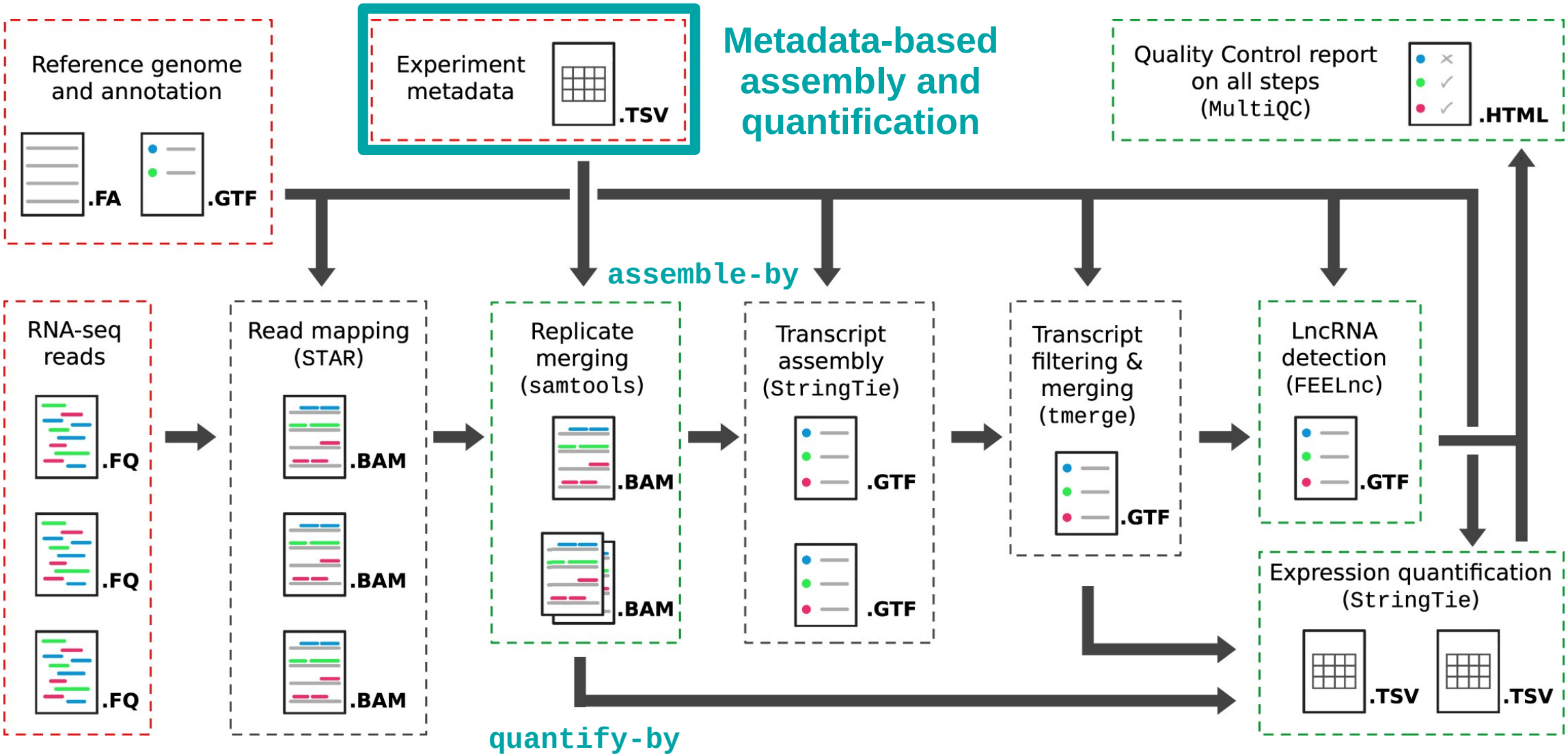
Published: 16 October 2023 **Article history**

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

The TAGADA pipeline



The TAGADA pipeline

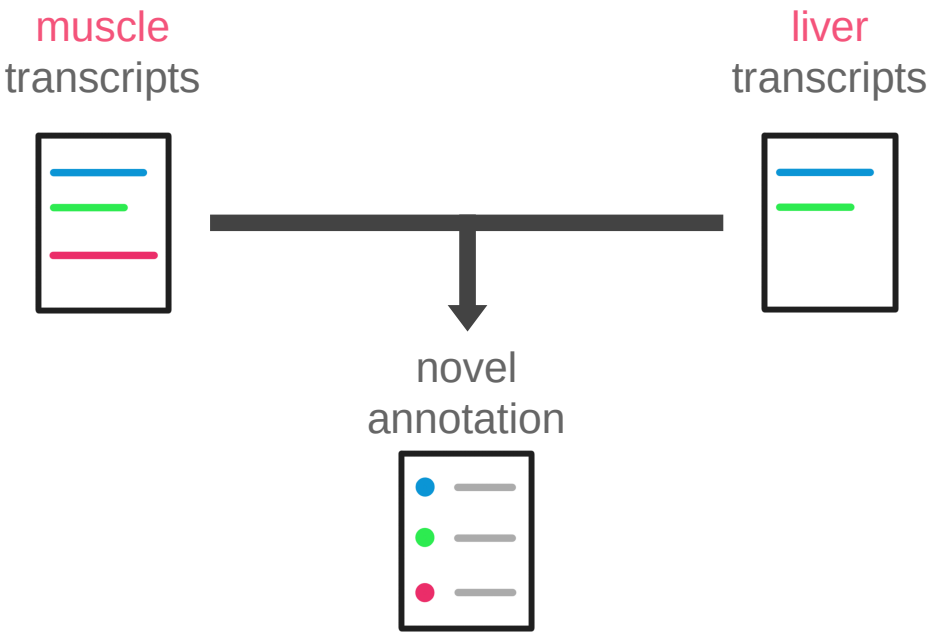


Simple and flexible replicate management

Metadata

name	tissue	day	breed
reads_A	muscle	70	Holstein
reads_B	muscle	30	Norwegian red
reads_C	muscle	30	Holstein
reads_D	liver	70	Norwegian red
reads_E	liver	70	Holstein
reads_F	liver	30	Holstein

Pipeline outputs



Command line option

```
--assemble-by   tissue
```

Simple and flexible replicate management

Metadata

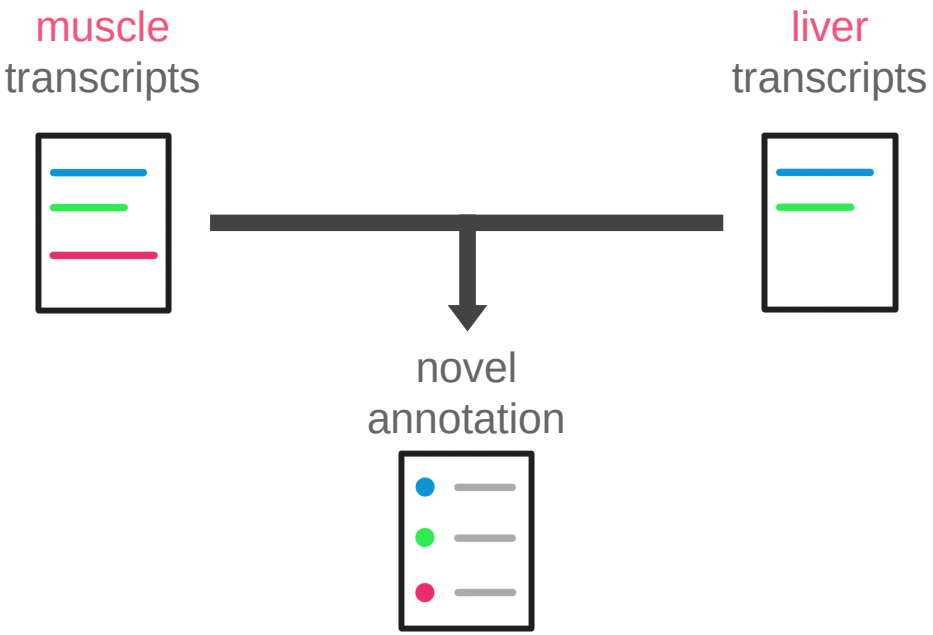
name	tissue	day	breed
reads_A	muscle	70	Holstein
reads_B	muscle	30	Norwegian red
reads_C	muscle	30	Holstein
reads_D	liver	70	Norwegian red
reads_E	liver	70	Holstein
reads_F	liver	30	Holstein

Command line option

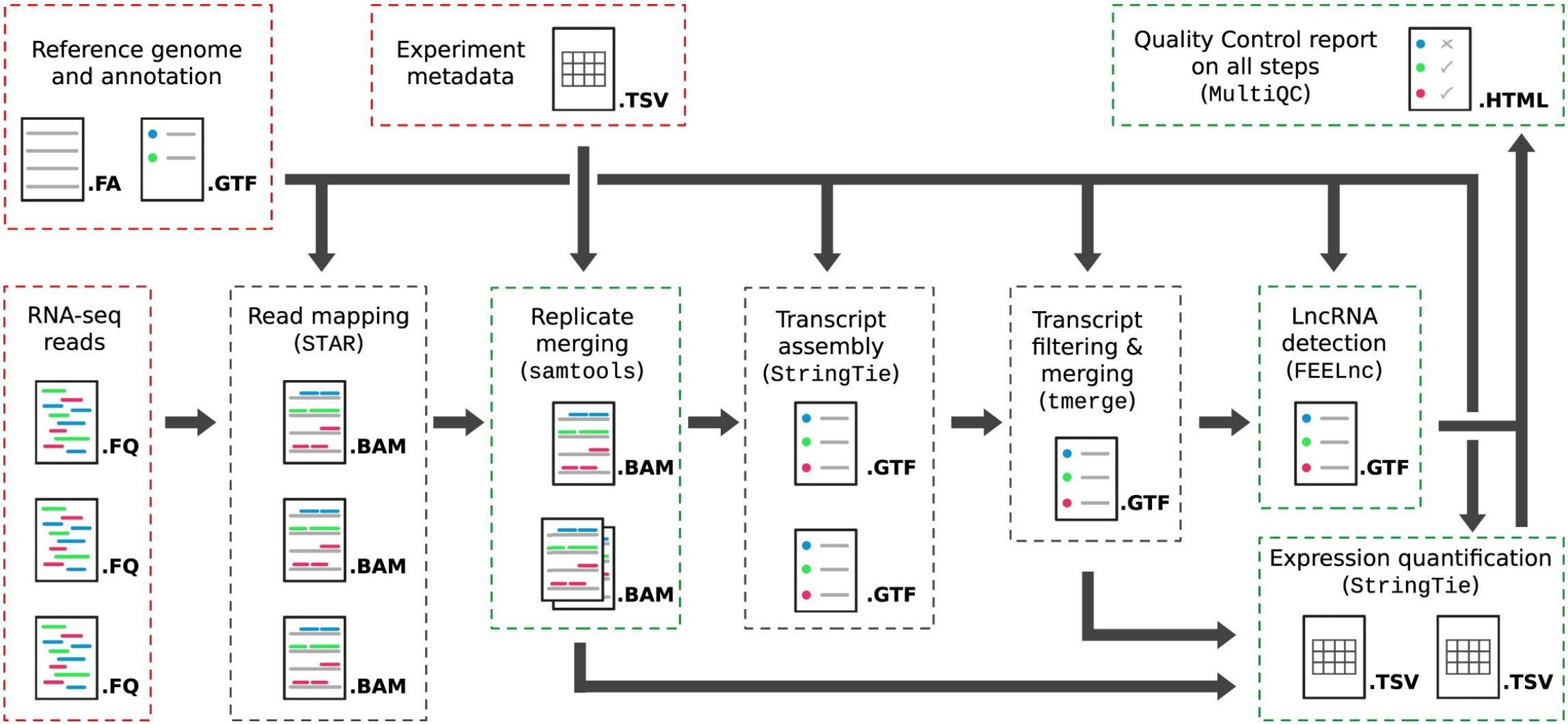
```
--assemble-by    tissue
```

```
--quantify-by    tissue,day
```

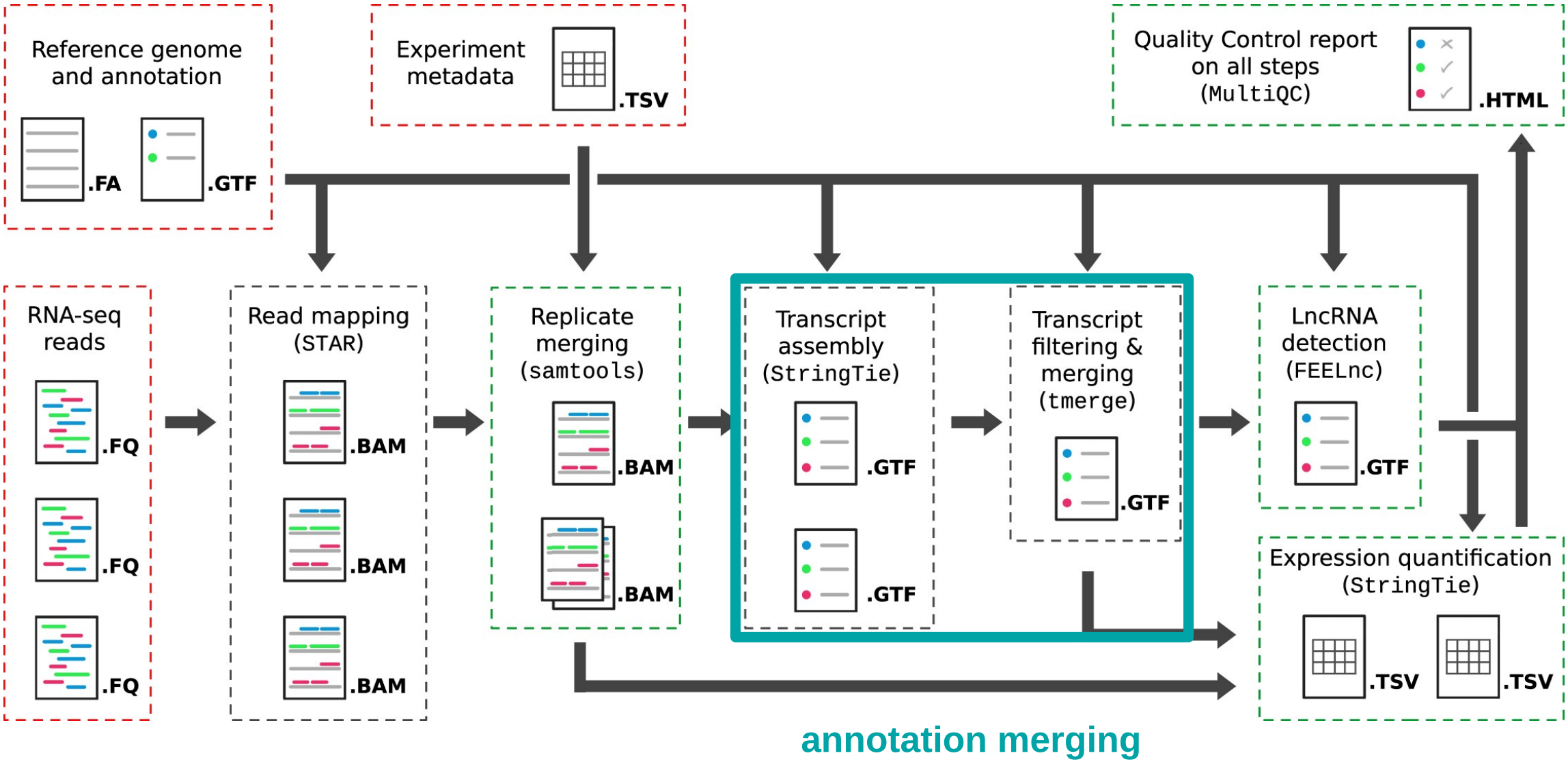
Pipeline outputs



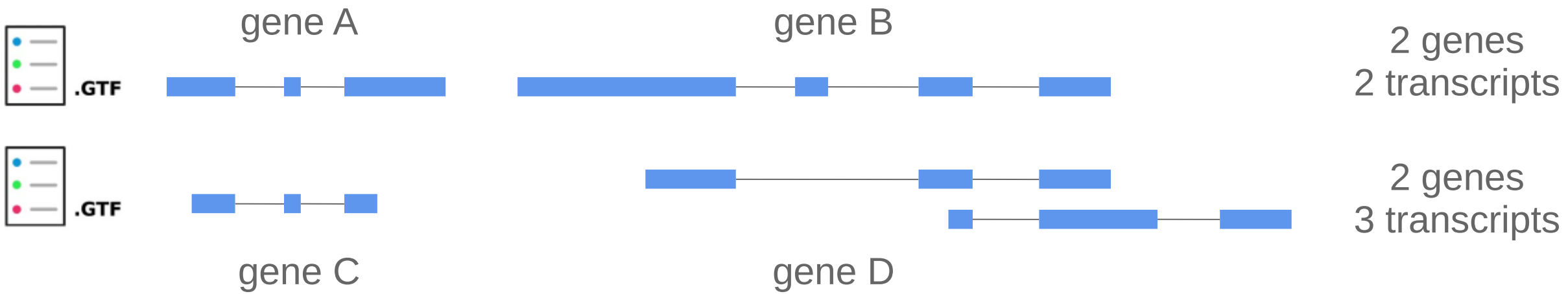
The TAGADA pipeline



The TAGADA pipeline

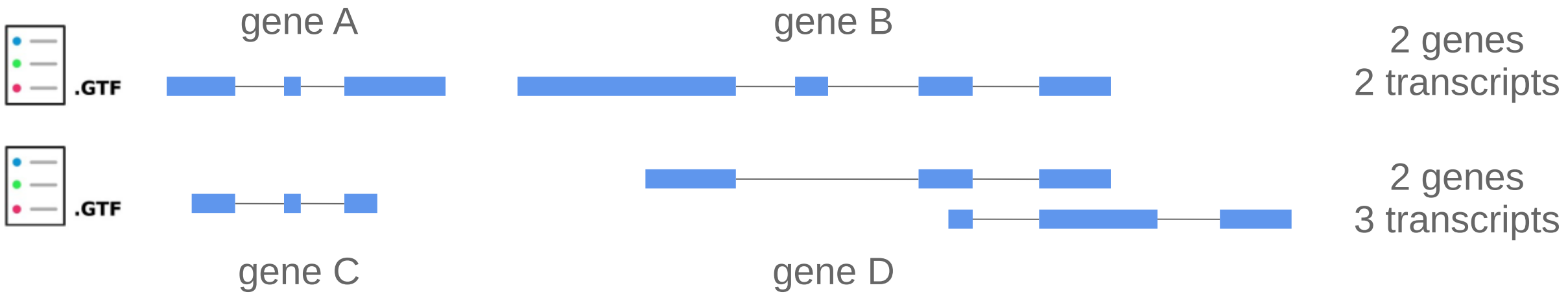


Merging annotations

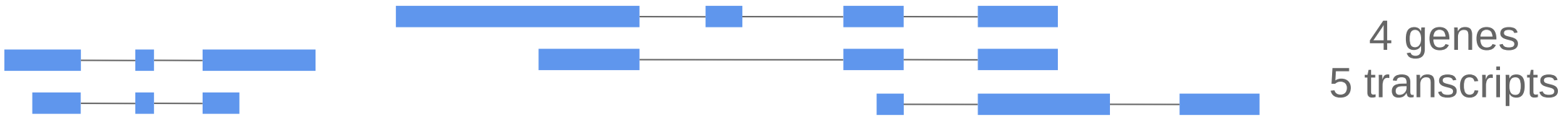


How to merge?

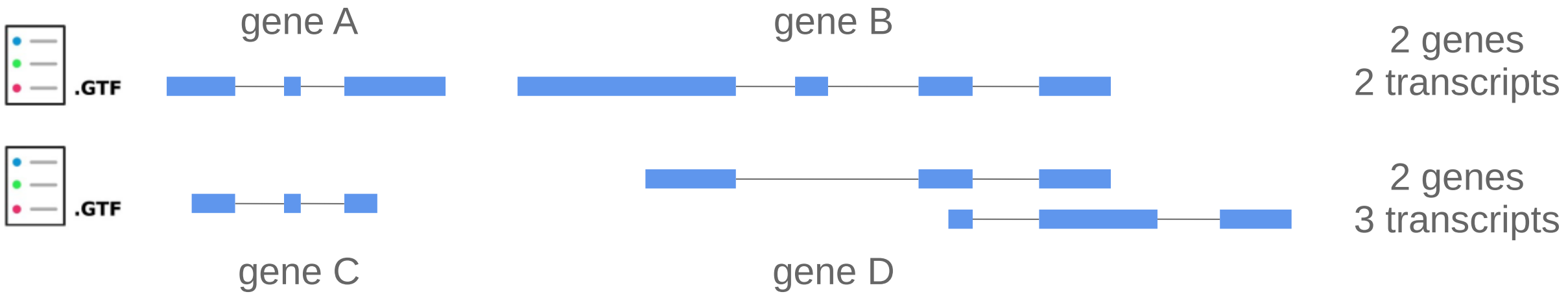
Merging annotations



Option “pile it up": concatenate everything

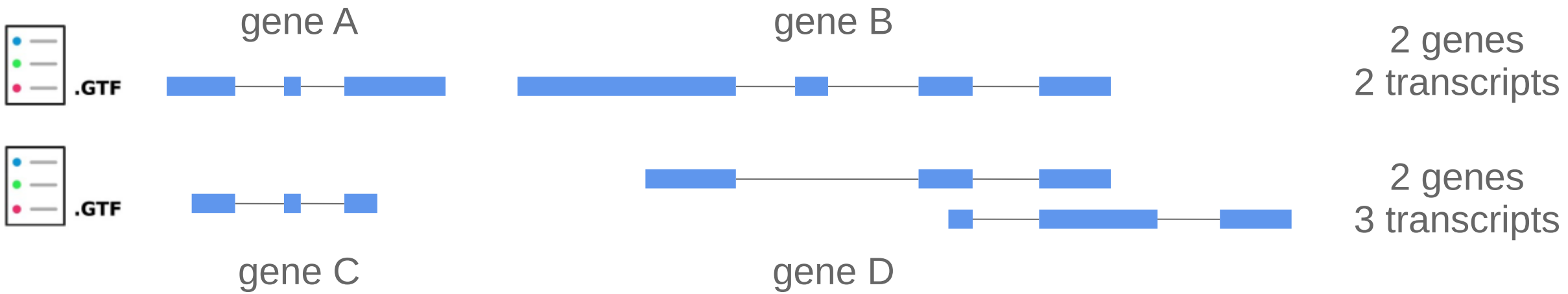


Merging annotations



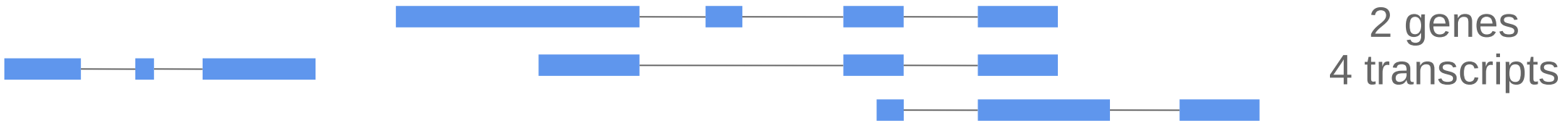
Option “contig it": merge all compatible structures

Merging annotations



Option “tmerge”: merge strict inclusions only

https://github.com/guigolab/tmerge



Additional features: expression-based filter, consistency-based filter, reference annotation inclusion

TAGADA implementation

README Apache-2.0 license

TAGADA: Transcript And Gene Assembly, Deconvolution, Analysis

TAGADA is a Nextflow pipeline that processes RNA-Seq data. It parallelizes multiple tasks quality, align reads to a reference genome, assemble new transcripts to create a novel a quantify genes and transcripts.

Table of contents

nextflow

Portable
Docker + Singularity

Scalable
Slurm, Kubernetes….

Modular
DSL2 processes

Reproducible

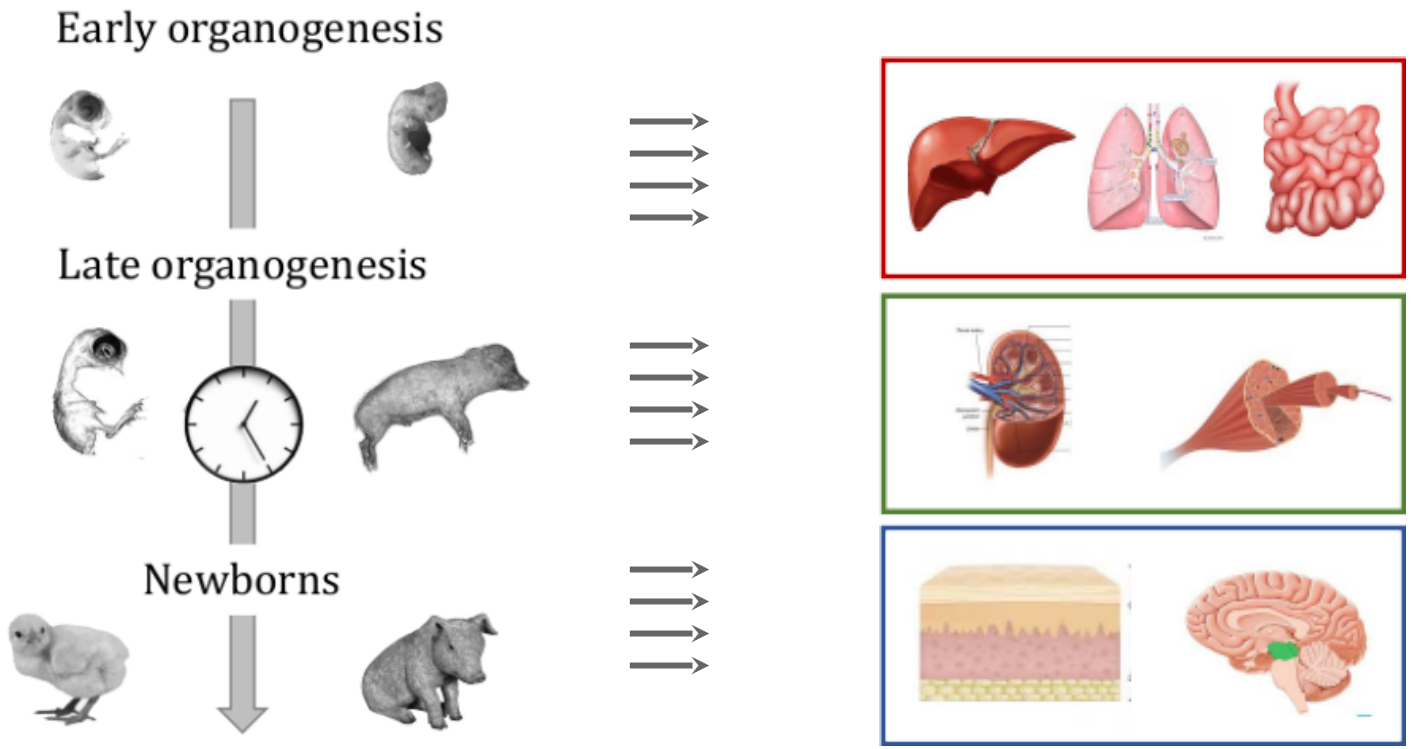
```
nextflow run FAANG/analysis-TAGADA \
     -profile test,docker \
     -revision 2.1.3 --output directory
```

Application example: the GENE-SWitCH project

- 2 species
  - *Sus scrofa* (pig)
  - *Gallus gallus* (chicken)
- 84 experiments per species
  - 3 developmental stages
  - 4 animals
  - 7 tissues
- PolyA+ RNA-seq (Illumina)
  - PE 2x150bp, directional
  - 100-150 million PE reads/library
- Reference genomes
  - chicken: Galgal6
  - pig: Sscrofa11.1
- Reference annotation: ensembl v102

GENE-SWitCH

3 dev. stages **4 replicates** **7 tissues**



2 species

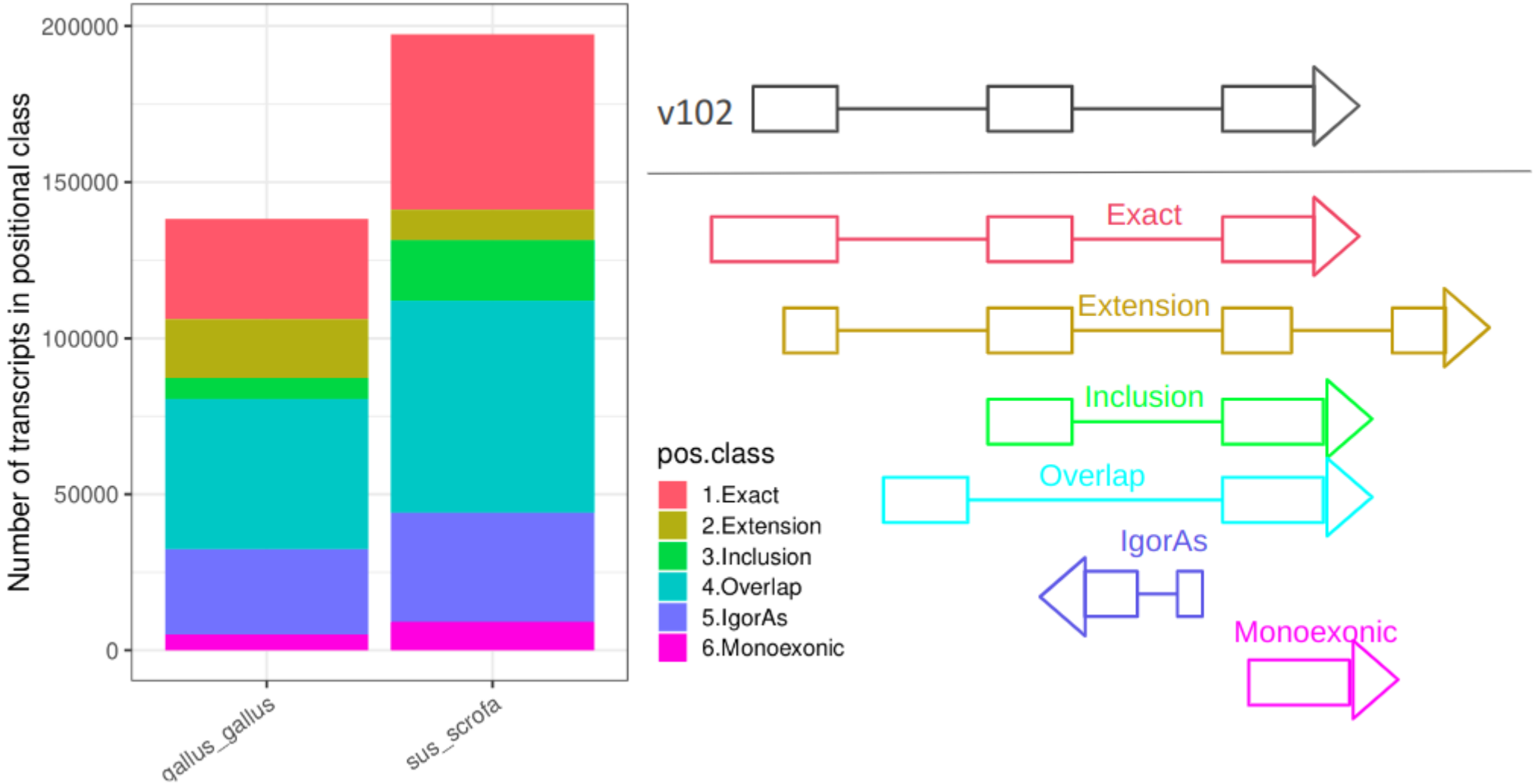
TAGADA annotation improves the reference

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

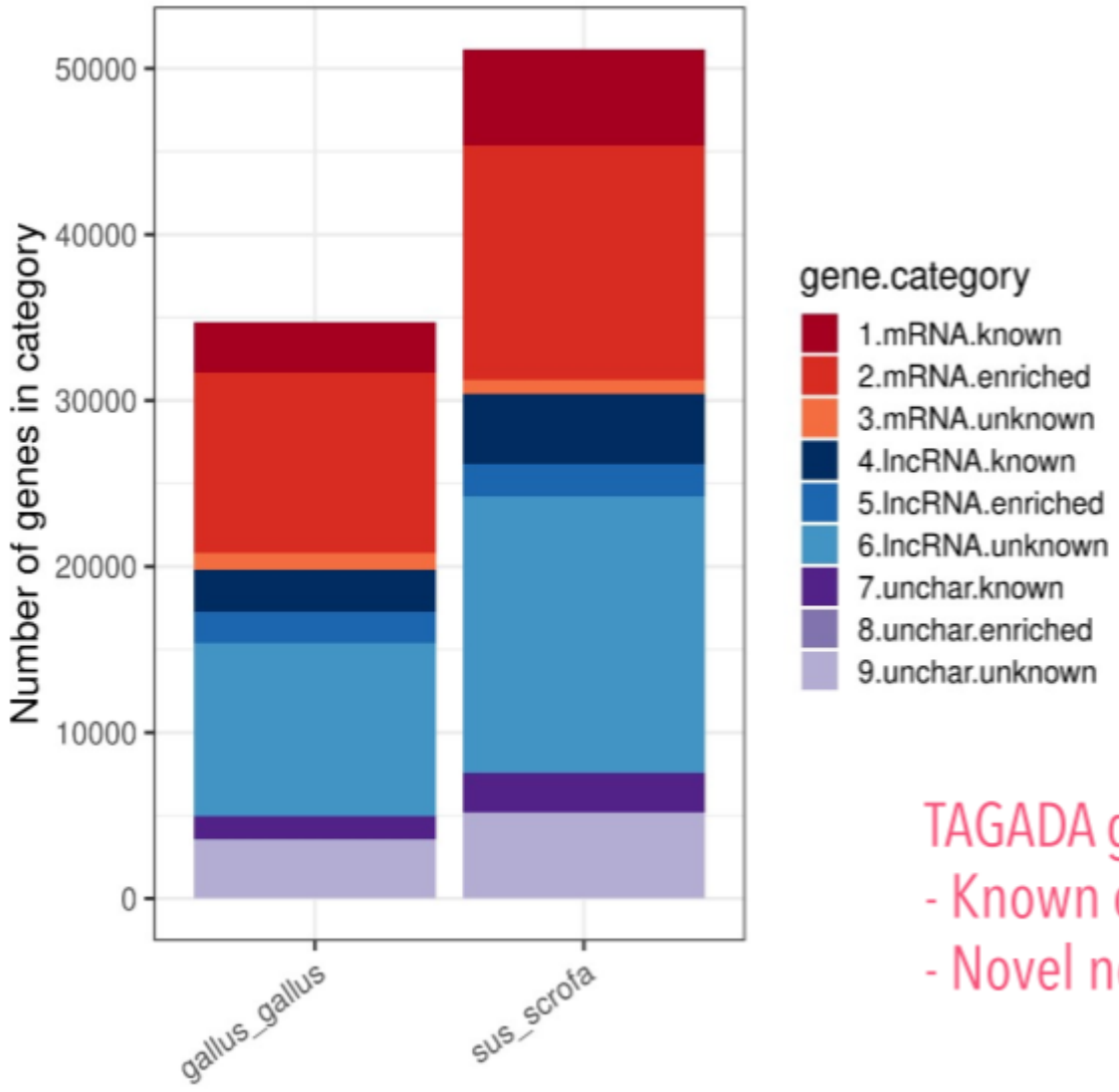
X 1.4-1.6 genes

x 3.1-3.5 transcripts

Characterizing the new annotation

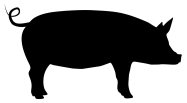


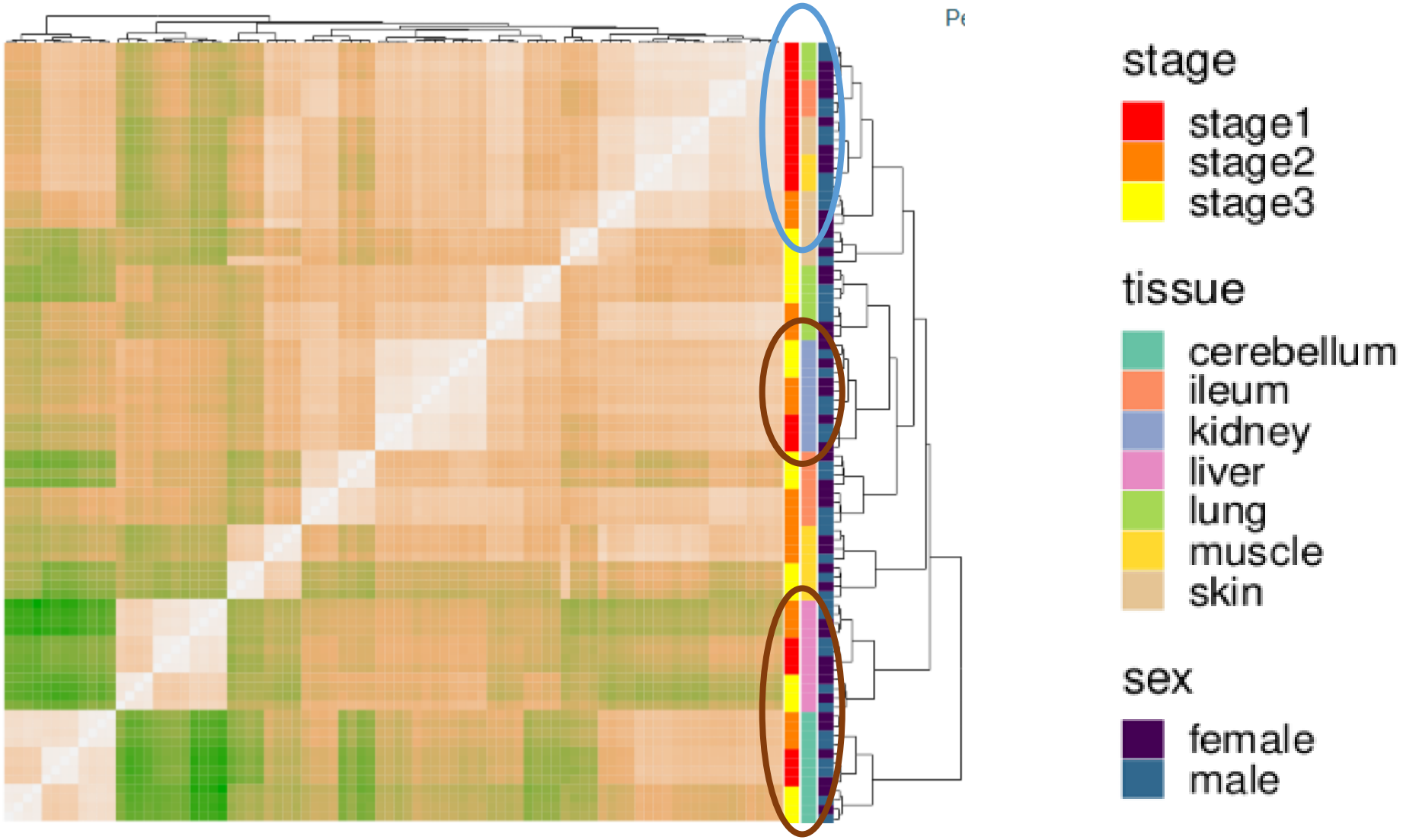
Characterizing the new annotation



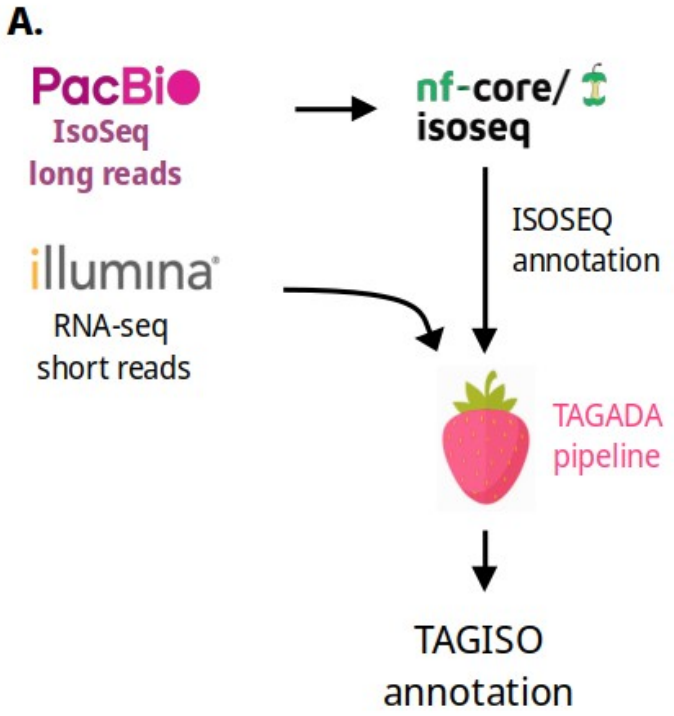
TAGADA genes are :
- Known coding genes with novel transcripts
- Novel non coding genes

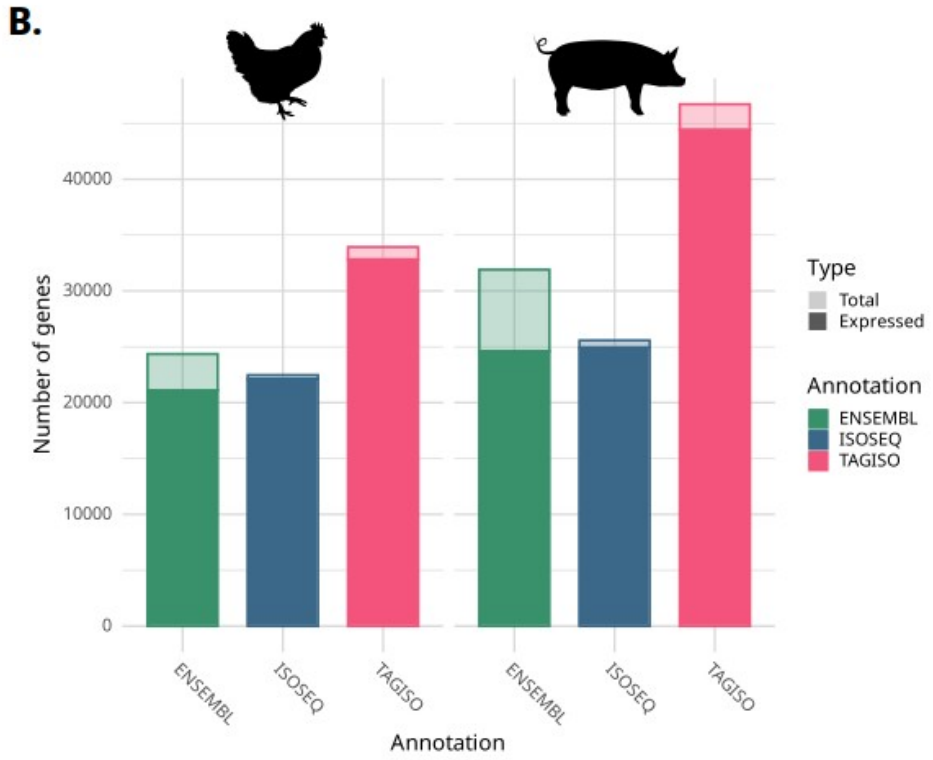
Expression-based sample clustering





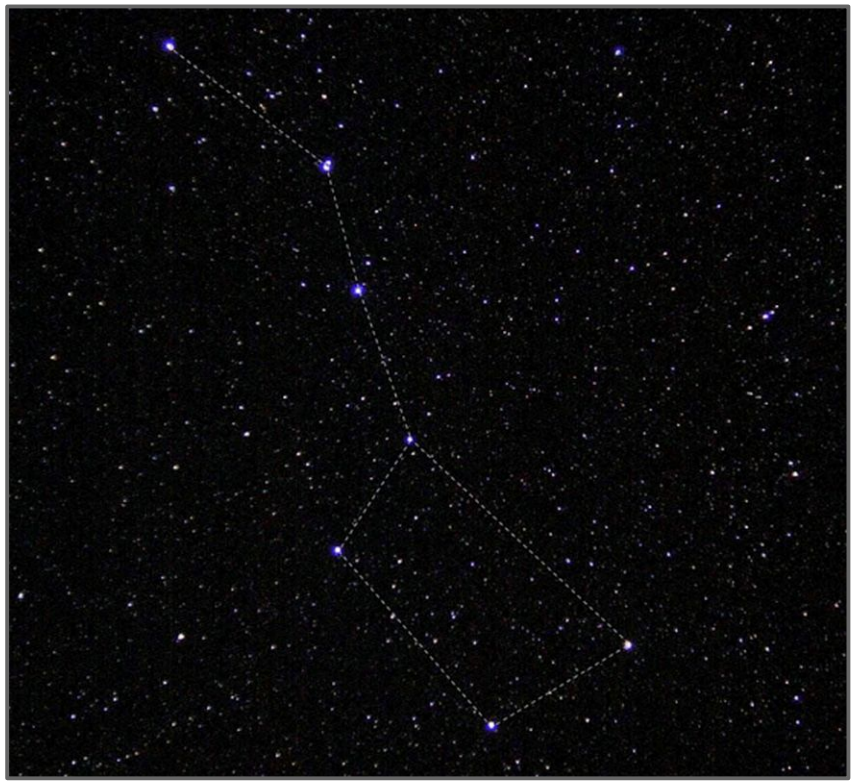
Combining PacBio and Illumina RNA-seq





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%**

Conclusion: so, what is a gene?

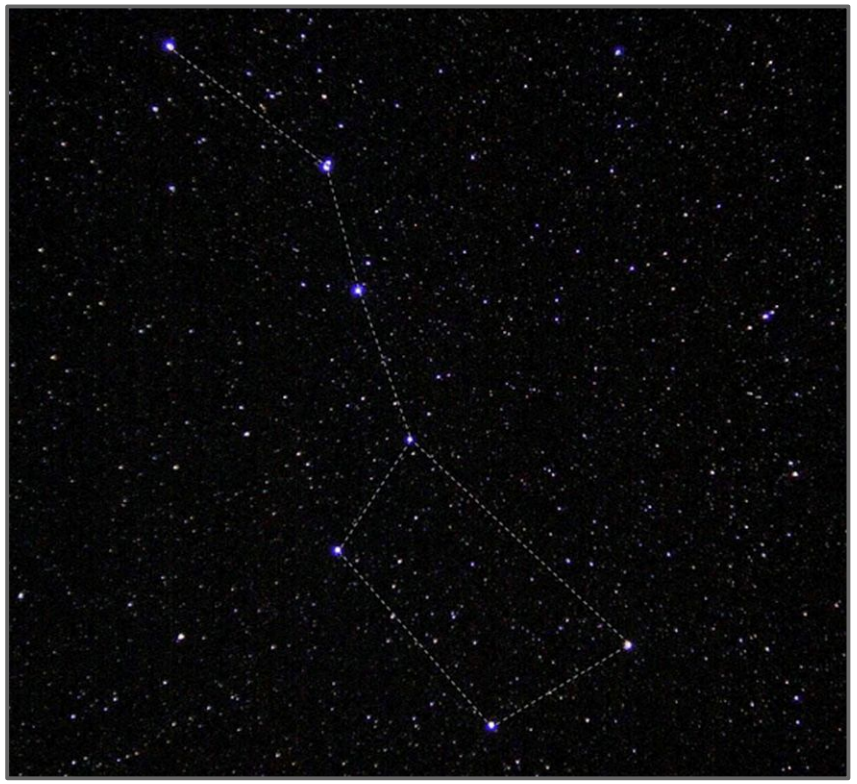


Conclusion: so, what is a gene?

the cloud view

- intangible
- innumerable
- no clear boundaries
- evanescent / transient / temporary

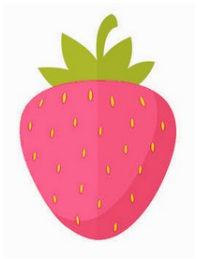
the constellation view



- group of identified entities
- artificial construction
- arbitrary names and structure
- important for navigation/exploration
- small fraction of the reality

Thank you!






Cyril Kurylo[1,†], Cervin Guyomar[1,†], Sylvain Foissac[1,*] and Sarah Djebali[2]

[1]GenPhySE, Université de Toulouse, INRAE, INPT, ENVT, Toulouse, France
[2]IRSD, Université de Toulouse, INSERM, INRAE, ENVT, Univ Toulouse III - Paul Sabatier (UPS), Toulouse, France

[*]To whom correspondence should be addressed. Email: sylvain.foissac@inrae.fr
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

This project has received funding from the European Union's ***Horizon 2020 Research and Innovation*** *Programme under the grant agreement* ***n° 817998***
@GeneSwitch - www.gene-switch.eu